Tutorial: Topological Data Analysis

Moo K. Chung

University of Wisconsin-Madison

https://github.com/laplcebeltrami/PH-STAT

mkchung@wisc.edu

Abstract

Topological data analysis (TDA) provides a mathematical framework for extracting multiscale geometric and topological structure from complex data. Unlike conventional methods designed for Euclidean data, TDA naturally analyzes point clouds, networks, images, and other non-Euclidean objects through persistent homology, graph filtrations, and related topological invariants. This tutorial introduces the statistical foundations of TDA, including simplicial complexes, filtrations, persistent homology, persistence diagrams, and topological distances, together with practical computation using the standalone open-source MATLAB toolbox PH-STAT (https://github.com/laplcebeltrami/PH-STAT). Hands-on examples will cover structural and functional brain connectomics, dynamic brain networks, and other non-Euclidean datasets.

The tutorial will also introduce recent advances in topological analysis of directed interactions and causal modeling, including Hodge-theoretic representations of feedback loops and cyclic information flow.

Participants will gain both the theoretical background and practical tools needed to apply topological methods to modern data analysis problems.

Lecture materials for TDA tutorial

Standalone TDA matlab toolbox (50+ functions)
https://github.com/laplcebeltrami/PH-STAT
Tutorial slides are in this site.
Plan to have 200+ functions

Detailed software manual with basic theory (49 pages). Plan to have 300+ page manual
https://arxiv.org/abs/2304.05912

Lecture slides have unlerlined links for additional information and papers.

Where topology came from?

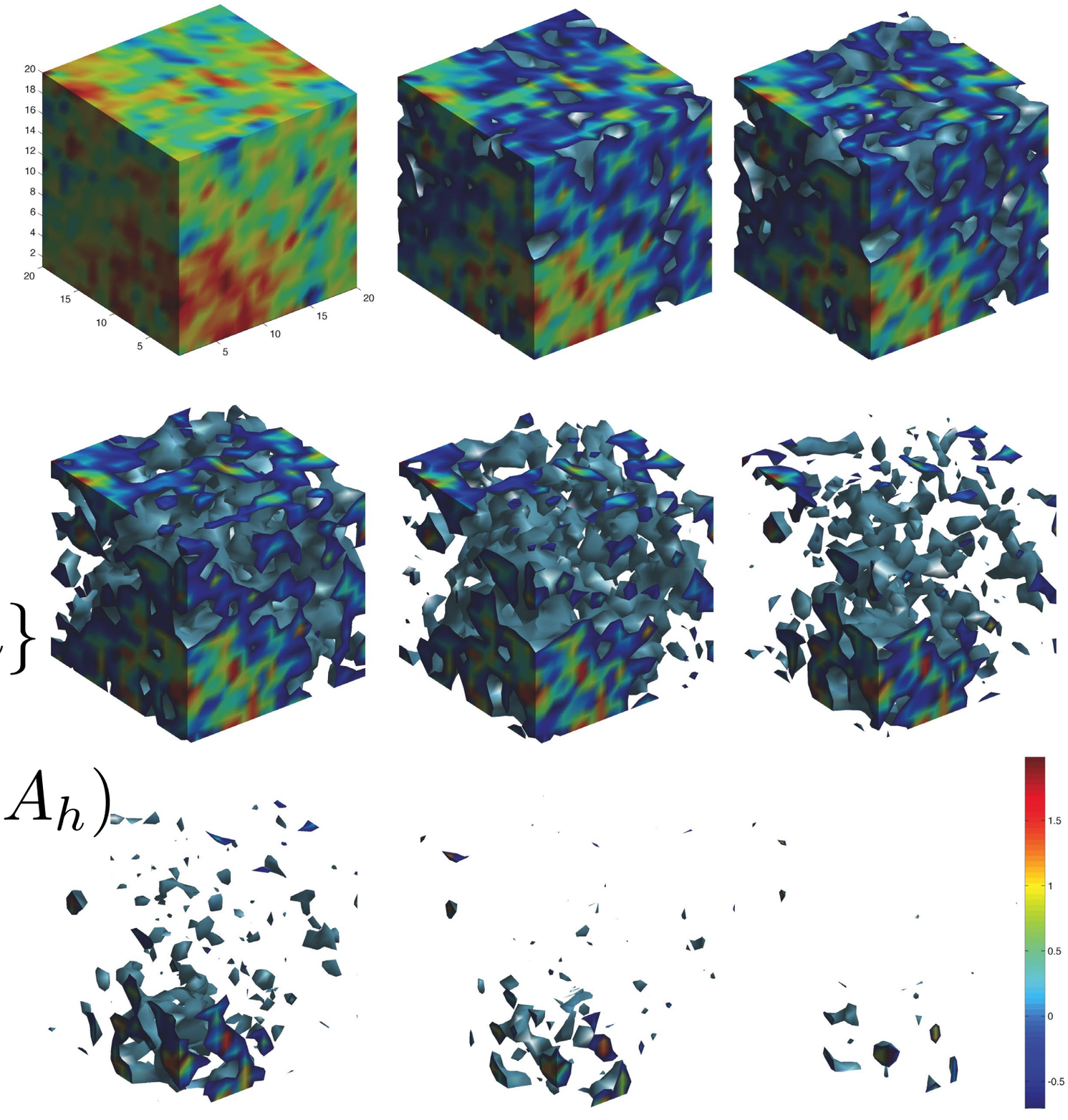
Random field

$$P\left(\sup_{x\in\mathbb{M}} T(x) > h\right)$$

$$A_h = \{x \in \mathbb{M} : T(x) > h\}$$

$$P\left(\sup_{x\in\mathbb{M}} T(x) > h\right) = \mathbb{E}\chi(A_h)$$

$$\chi(A_h) = \sum_j (-1)^j \beta_j(A_h)$$



Milor 1963 Morse theory

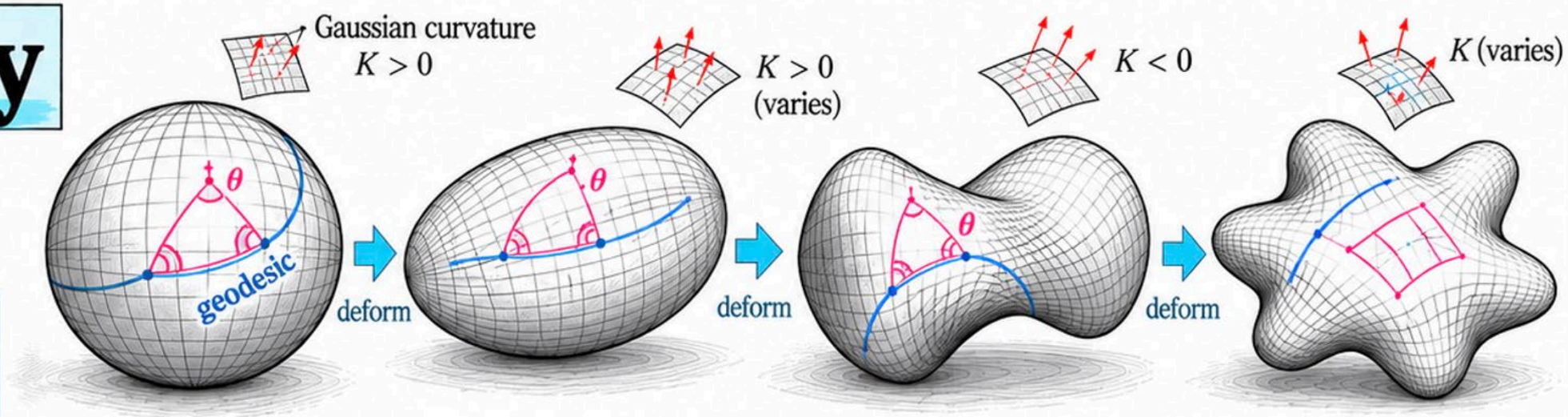
Adler, 1994 The geometry of random field

Worsley et al., 1996 Human Brain Mappin

Geometry

Quantities that **change** under continuous deformation.

- length (geodesic distance)
- angle
- curvature
- area / surface area
- volume
- metric, coordinates, …



Sphere
Area $4\pi R^2$
Volume $\frac{4}{3}\pi R^3$
$K = 1/R^2 > 0$
Sum of angles (triangle) $> \pi$

Ellipsoid (stretched)
Area $> 4\pi R^2$
Volume $= \frac{4}{3}\pi R^3$
$K > 0$ (varies)
Sum of angles $> \pi$

Saddle surface
Area (varies)
Volume (varies)
$K < 0$
Sum of angles $< \pi$

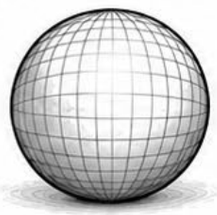
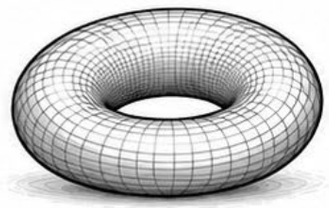
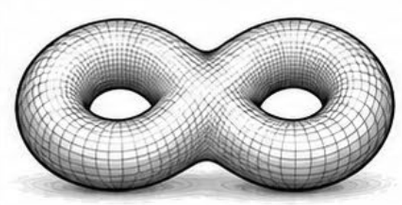
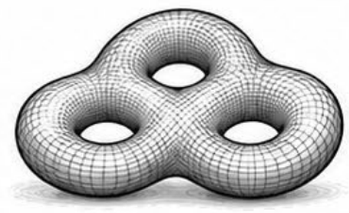
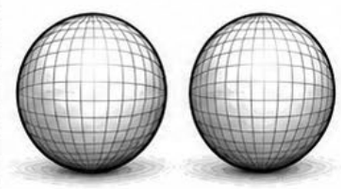
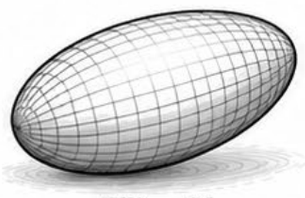
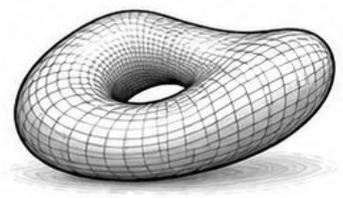
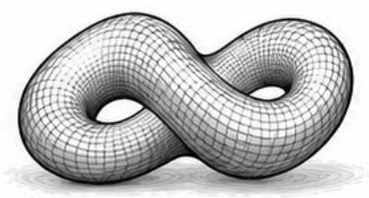
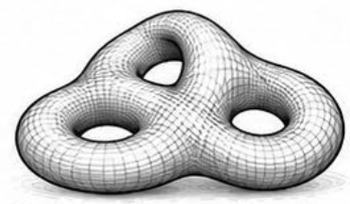
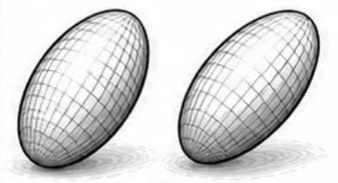
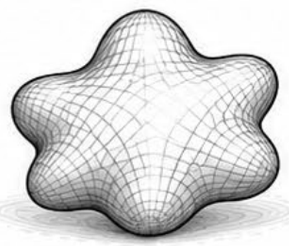
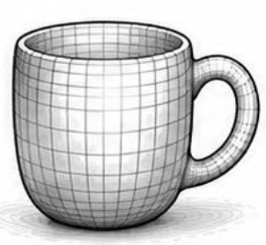
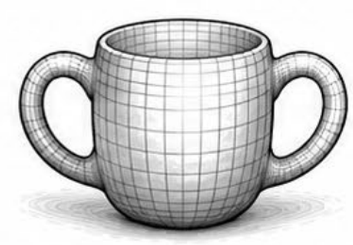
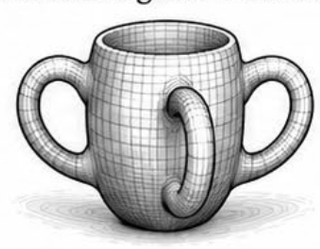
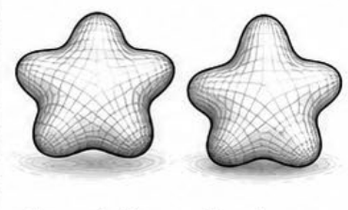
Deformed surface
Area (varies)
Volume (varies)
K (varies, $+/-$)
Sum of angles (varies)

Topology

Quantities that **remain invariant** under continuous deformation.

- number of connected components β_0
- number of holes (1D) β_1
- number of voids (2D) β_2
- genus g
- cannot change without cutting or gluing

Shapes in the same column are topologically equivalent (they can be continuously deformed into each other).

Genus 0 (sphere topology)	**Genus 1** (torus topology)	**Genus 2** (double-torus topology)	**Genus 3** (triple-torus topology)	**Two components** (disconnected surfaces)
$\beta_0 = 1,\ \beta_1 = 0,\ \beta_2 = 1$ $g = 0$	$\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 1$ $g = 1$	$\beta_0 = 1,\ \beta_1 = 4,\ \beta_2 = 1$ $g = 2$	$\beta_0 = 1,\ \beta_1 = 6,\ \beta_2 = 1$ $g = 3$	$\beta_0 = 2,\ \beta_1 = 0,\ \beta_2 = 2$ $g = 0$ (each)
 Sphere	 Torus	 Double torus	 Triple torus	 Two spheres
 Ellipsoid	 Deformed torus	 Deformed double torus	 Deformed genus-3 surface	 Two ellipsoids
 Deformed sphere	 Mug (handle)	 Two-handle mug	 Three-handle mug	 Two deformed spheres

Homology studies which objects cannot be continuously contracted to a lower-dimensional object (ultimately a point).

Betti number $\beta_0 = dim(H_0)$

Connected components. Every connected component contracts to a point.

counts the resulting points.

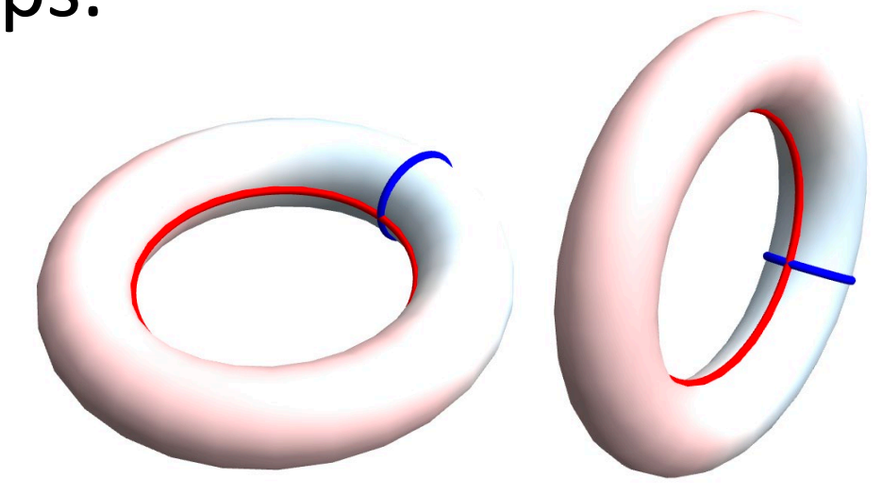


$\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$

Loops that cannot be contracted to a point

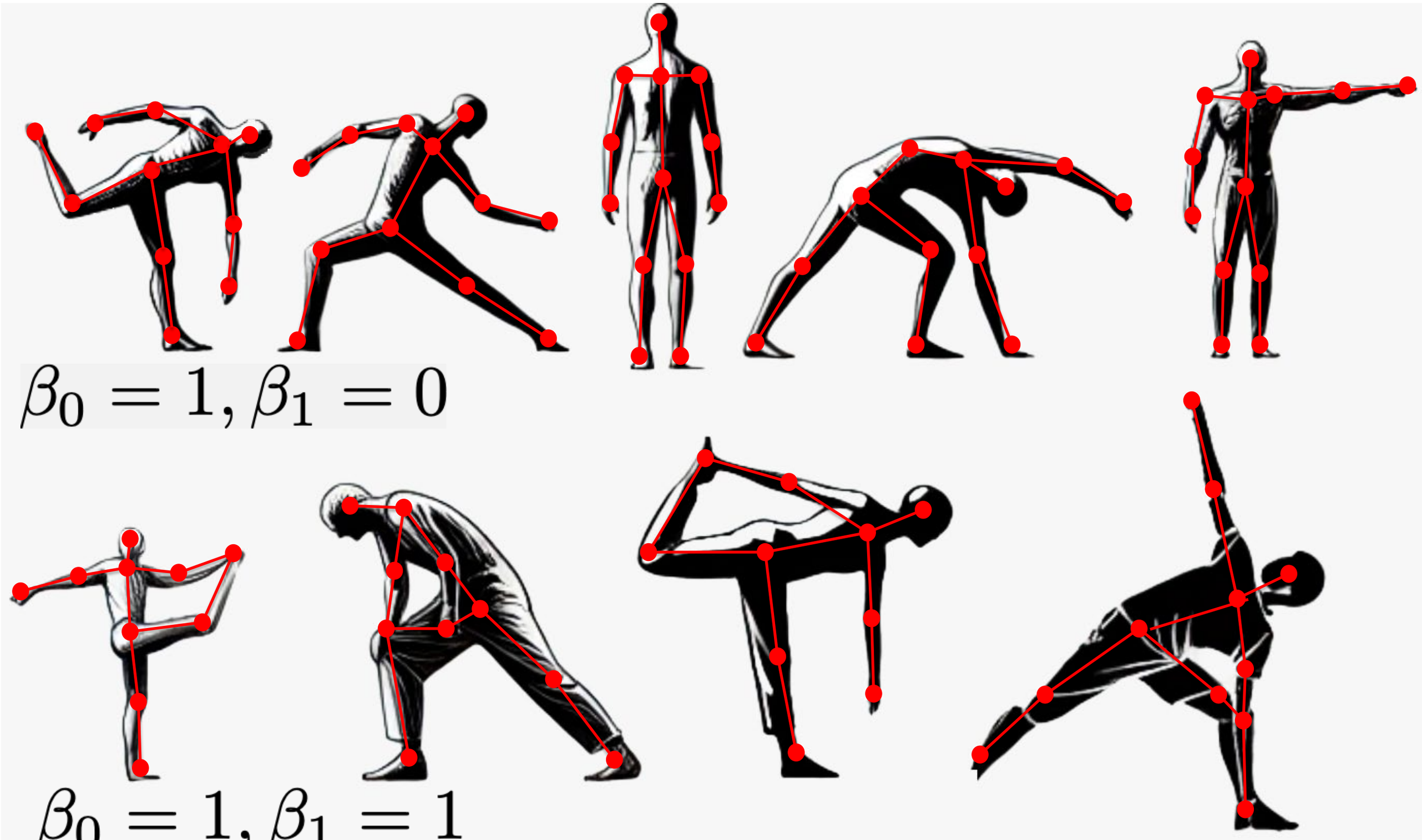
$\beta_1 = dim(H_1)$ counts the resulting loops.

H_2 Closed surfaces that cannot be contracted to a point because they enclose a cavity.



$\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$

Topological variability is zero.



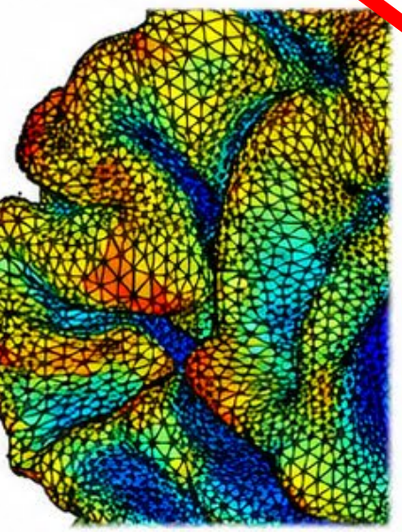
$\beta_0 = 1, \beta_1 = 0$

$\beta_0 = 1, \beta_1 = 1$

Can you come up with a method that gives 0 variability?

Morse filtration on functional data

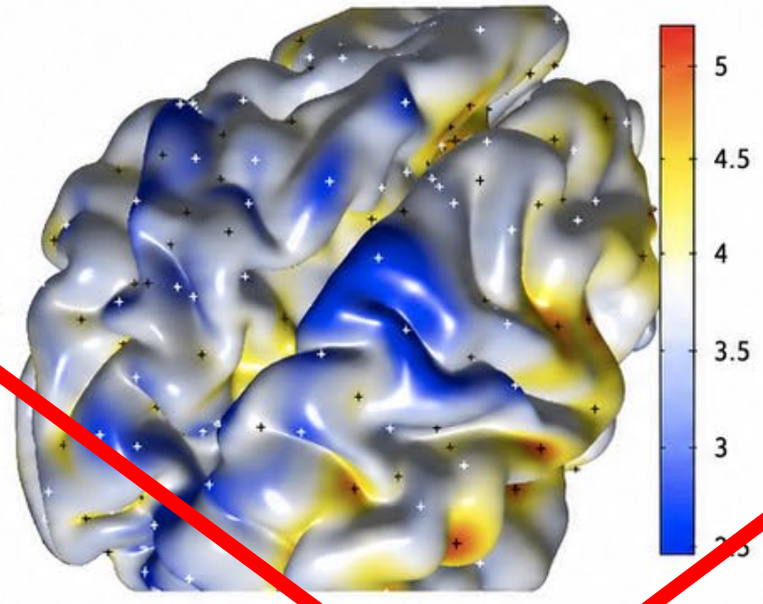
Functional data
on cortical surface $\mathcal{M}$



$f : \mathcal{M} \to \mathbb{R}$

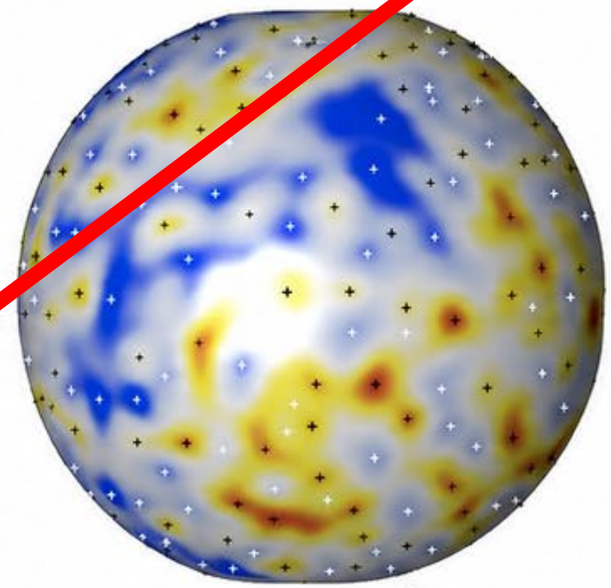
Diffusion

$u_t = e^{-t\Delta} f$



Flattening

$\phi : \mathcal{M} \to \mathbb{R}^2$

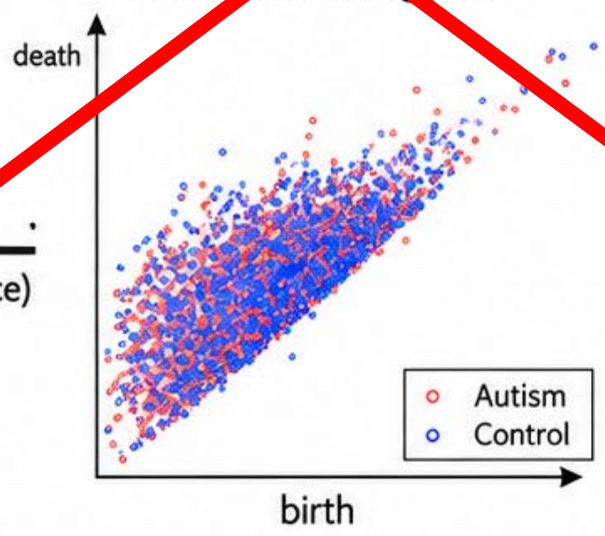


Delaunay
Triangulation

Morse function on mesh

$f_K : K \to \mathbb{R}$

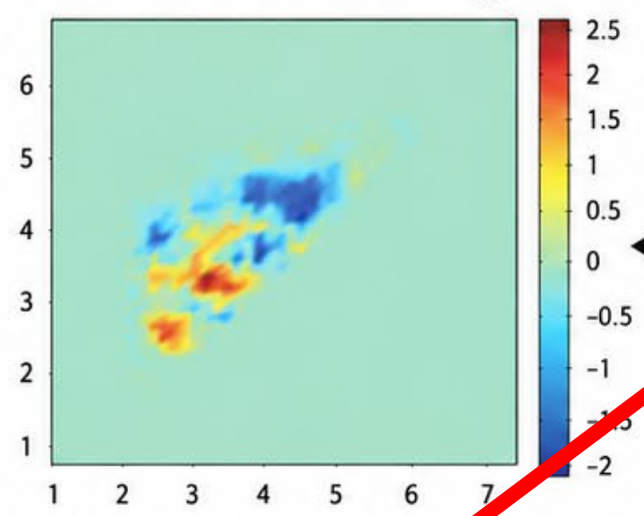
Persistent
Diagram



Smoothing

(in diagram space)

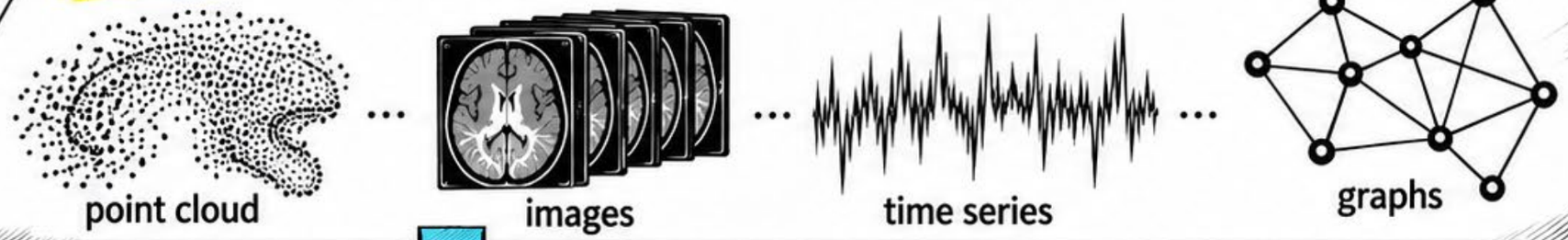
Smoothed Persistent Diagram



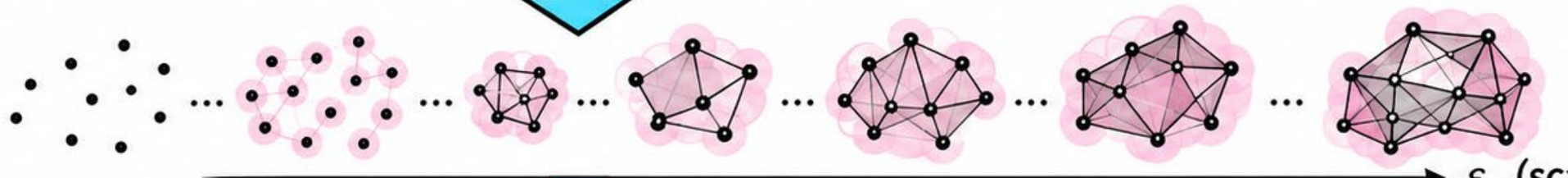
Chung et al., 2009

Distance-based TDA workflow

Data: functional, point cloud, images, graphs



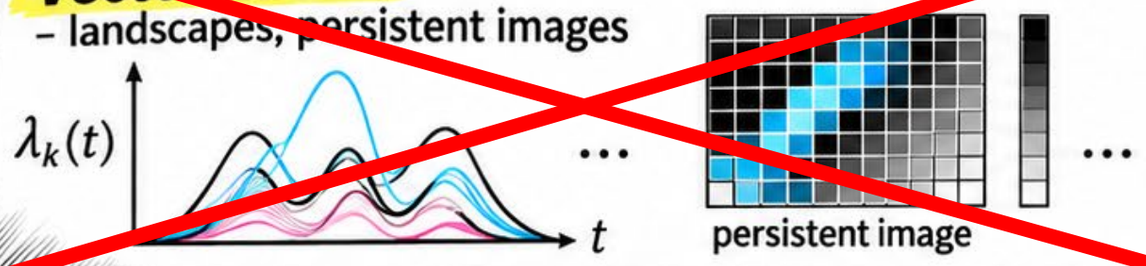
Filtration:
nested sets
→ multiscale, hierarchy



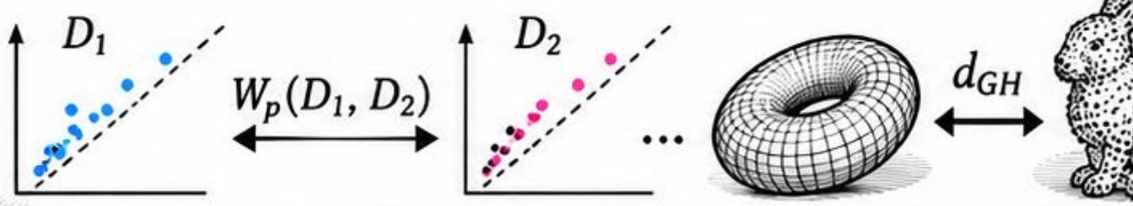
Topological features:
persistent diagrams, barcodes



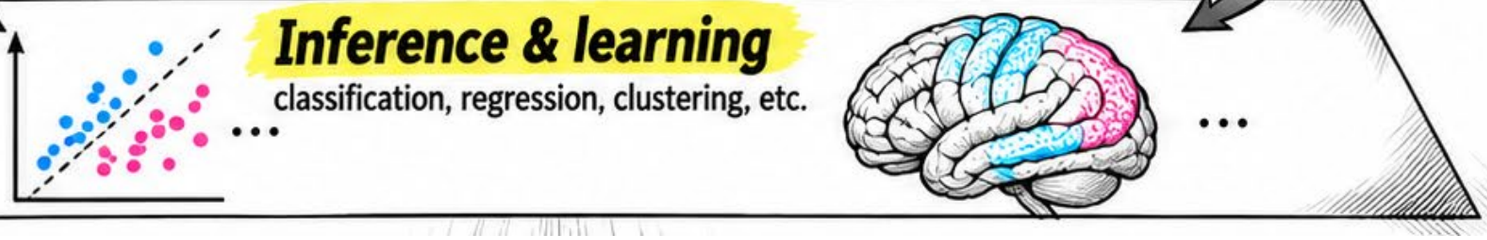
Vectorization
– landscapes, persistent images



Topological distance
Wasserstein distance, bottleneck, Gromov–Hausdorff



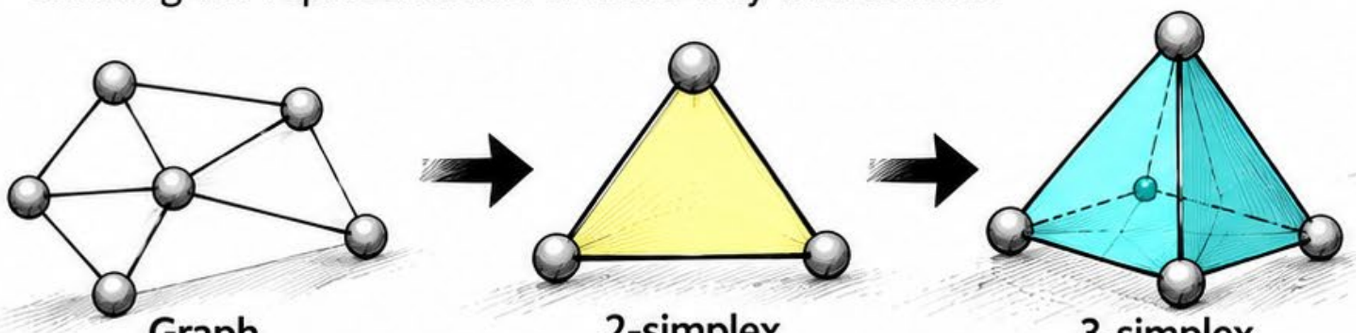
Inference & learning
classification, regression, clustering, etc.



Songdechakraiwut and Chung 2023 Annals of Applied

Simplicial complex Data Structure

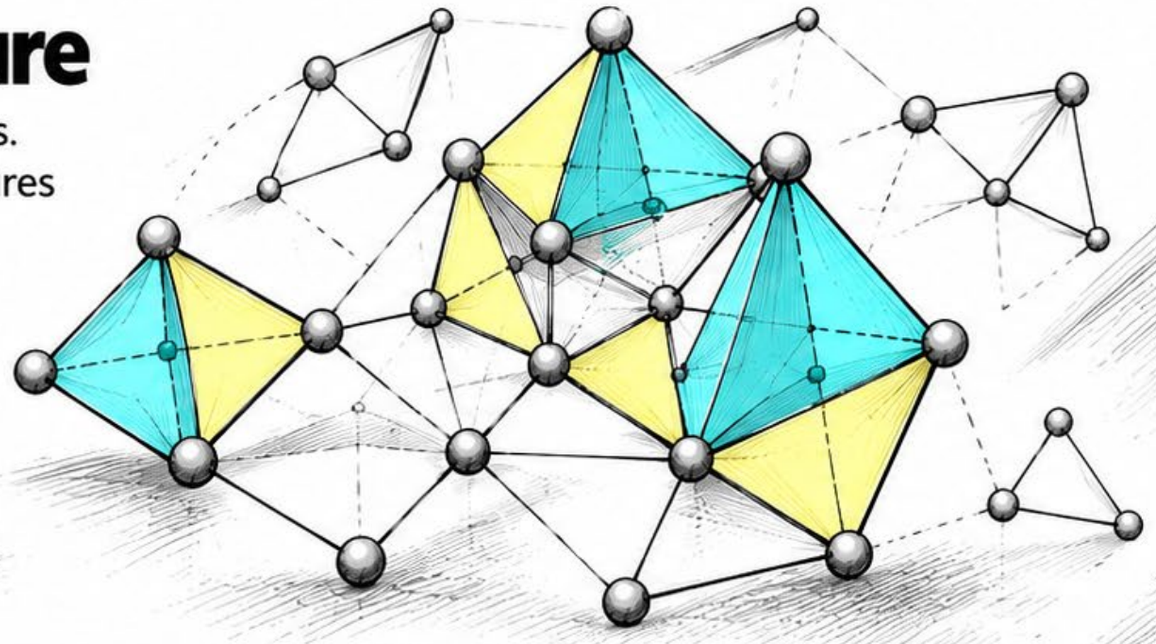
A **graph** represents only pairwise interactions through vertices and edges.
A **simplicial complex** extends a graph by including higher-order structures such as filled triangles (2-simplices) and filled tetrahedra (3-simplices), enabling the representation of multi-way interactions.



Graph
(vertices + edges)

2-simplex
(filled triangle)

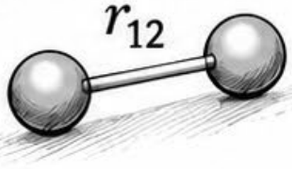
3-simplex
(filled tetrahedron)



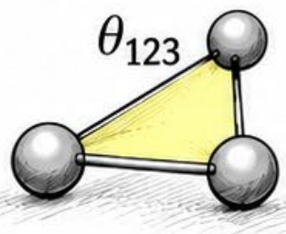
Simplicial complex

Why simplicial complex?

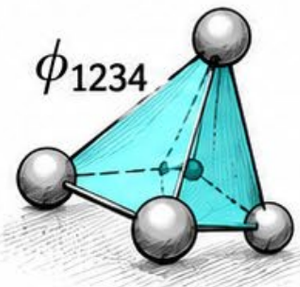
Encoding **higher order interactions** in protein structure



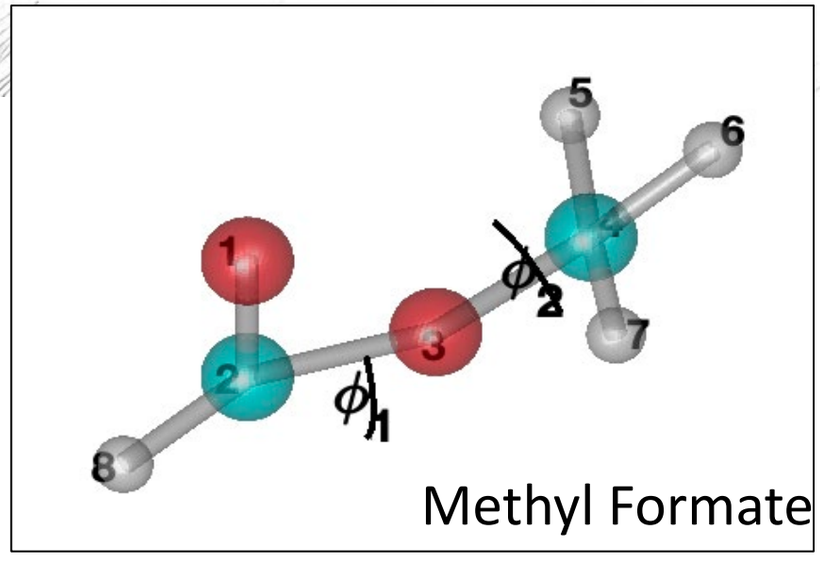
2-atom interaction



3-atom interaction
(3 atoms form a plane)



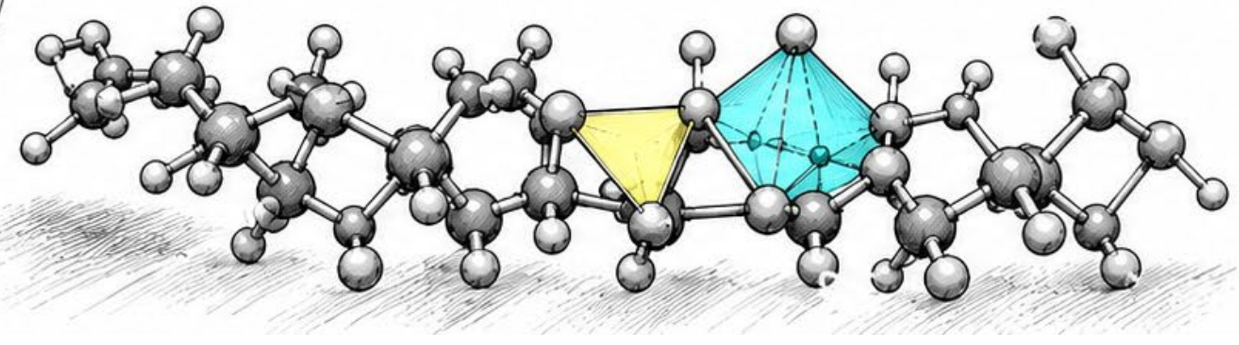
4-atom interaction
(dihedral angle)



Methyl Formate

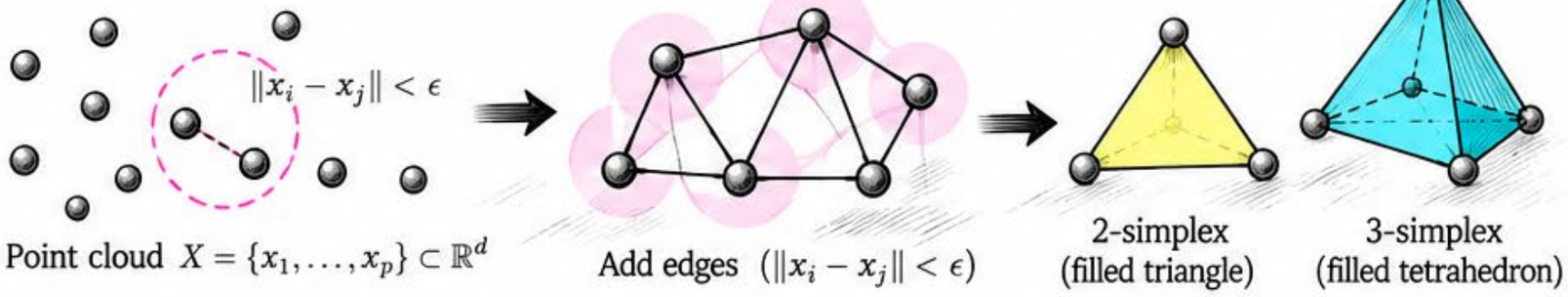
Potential energy

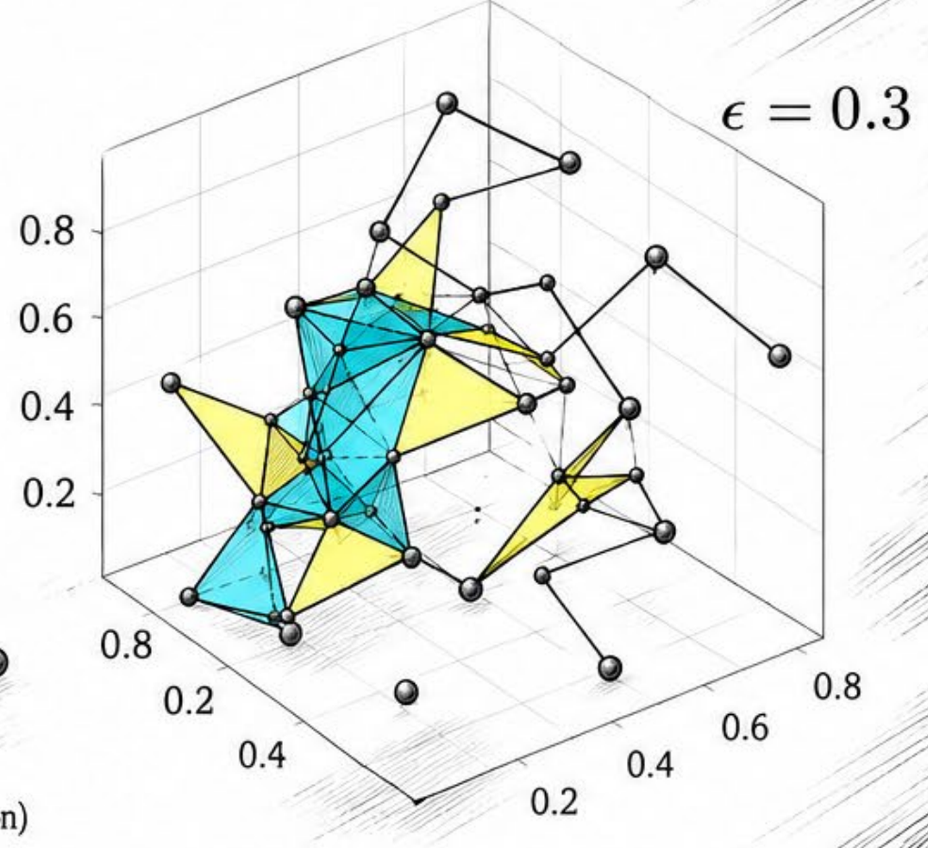
$$V = \underbrace{\sum_{(i,j)\in B} V_2(r_{ij})}_{\text{function over edges}} + \underbrace{\sum_{(i,j,k)\in A} V_3(\theta_{ijk})}_{\text{function over triangles}} + \underbrace{\sum_{(i,j,k,l)\in D} V_4(\phi_{ijkl})}_{\text{function over tetrahedra}}$$



Rips complex

Baseline simplicial complex is the **Rips complex** constructed by connecting pairs of points whose distance is less than radius ϵ. Whenever all edges of a simplex are present, its interior is filled automatically.



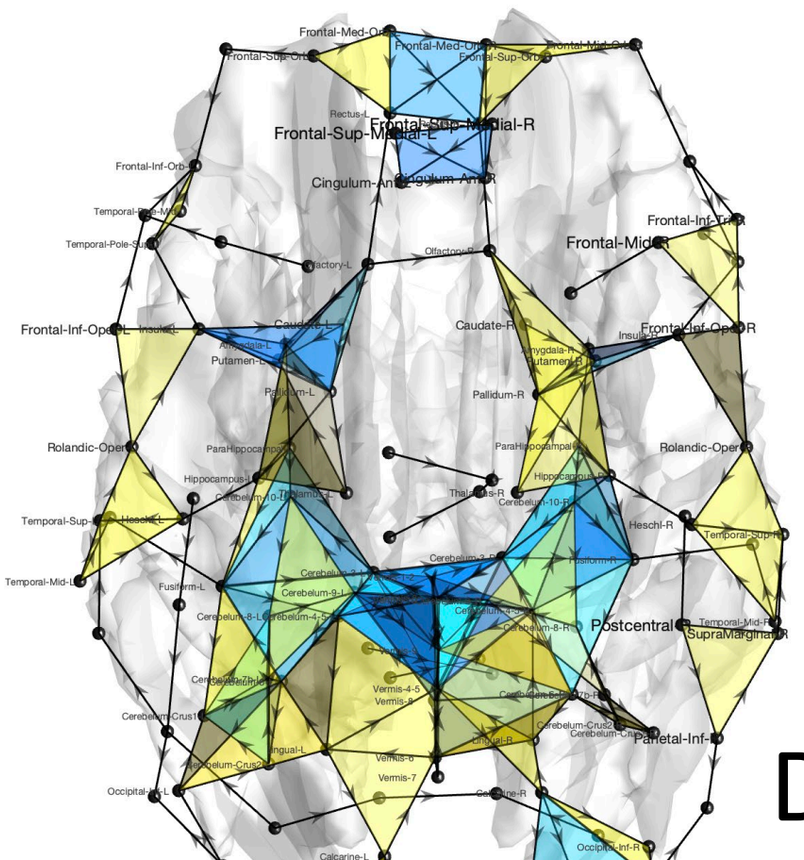


Matlab implementation in PH-STAT

```
p = 50; d = 3;
X = rand(p, d);

k = 3;
radius = 0.3;
S = PH_rips(X, k, radius);
PH_rips_display(X, S);
```

```
S = 4×2 cell array
  {50×1 double}    % vertices (0-simplices)
  {101×2 double}   % edges (1-simplices)
  {73×3 double}    % triangles (2-simplices)
  {27×4 double}    % tetrahedra (3-simplices)
```



Connectivity

(triangles)

$$\begin{bmatrix} 1 & 5 & 13 \\ 1 & 5 & 16 \\ 1 & 5 & 26 \\ \cdots & & \end{bmatrix}$$

(tetrahedra)

$$\begin{bmatrix} 2 & 7 & 18 & 24 \\ 3 & 5 & 12 & 20 \\ 3 & 6 & 14 & 28 \\ \cdot & & & \\ \cdots & & & \end{bmatrix}$$

Functional data over simplex

$$\begin{bmatrix} 0.2932 \\ 0.2579 \\ 0.2766 \\ \cdots \end{bmatrix}$$

$$\begin{bmatrix} 0.1254 \\ 0.3187 \\ 0.2191 \\ \cdots \end{bmatrix}$$

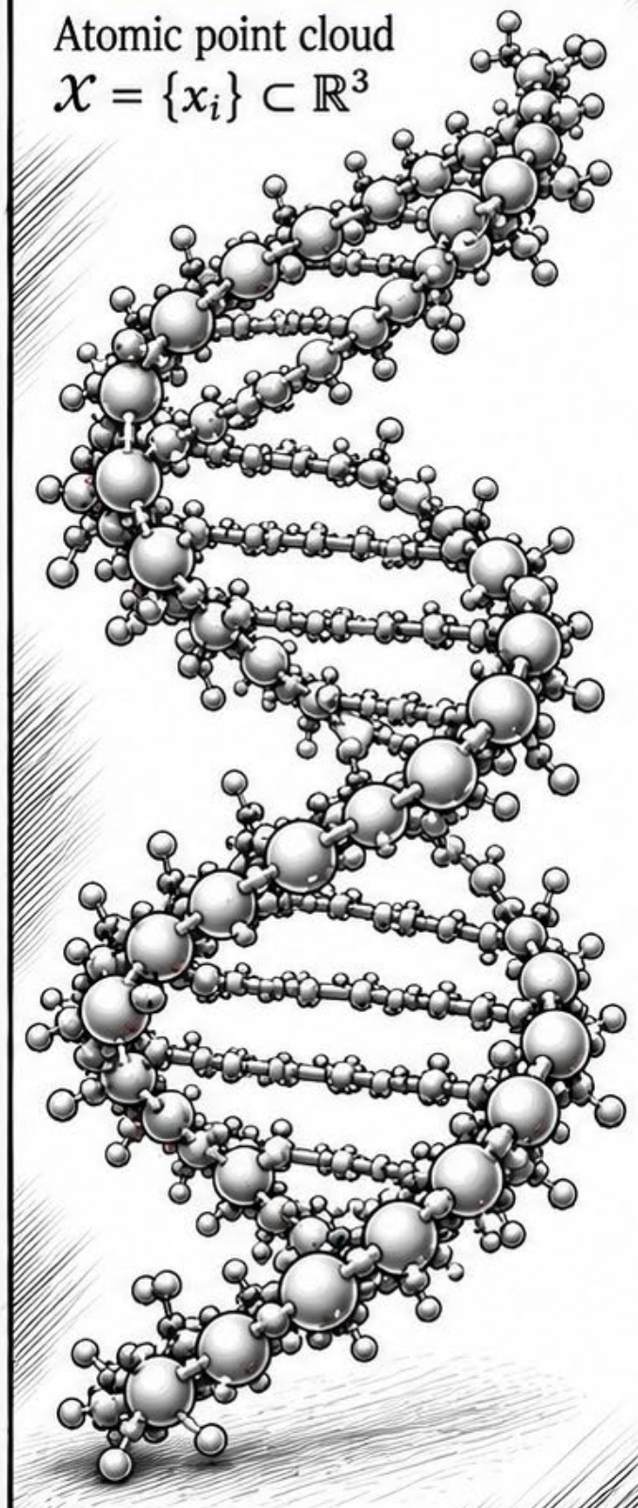
Directed interactions

Rips filtration

DNA double helix

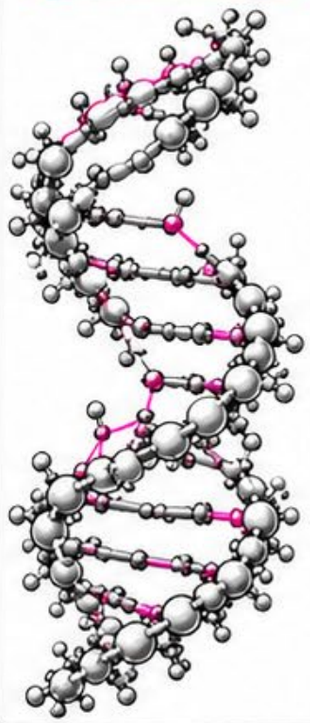
Atomic point cloud

$\mathcal{X} = \{x_i\} \subset \mathbb{R}^3$

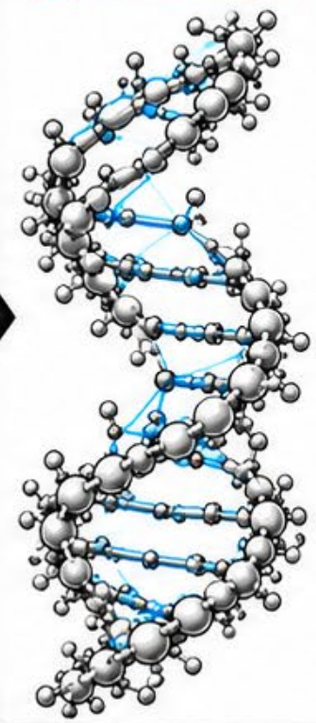


Rips filtration: connect points within distance ε

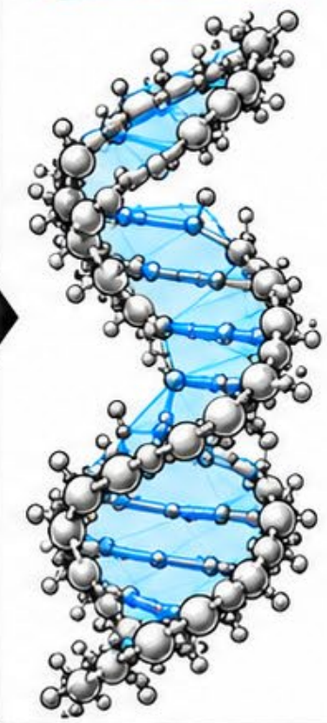
$\varepsilon = 1.5$ Å



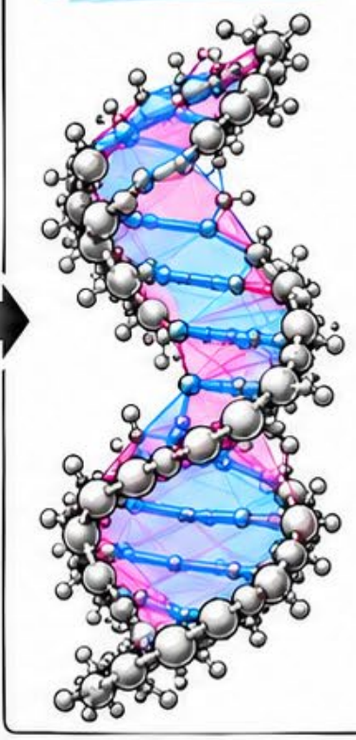
$\varepsilon = 2.5$ Å



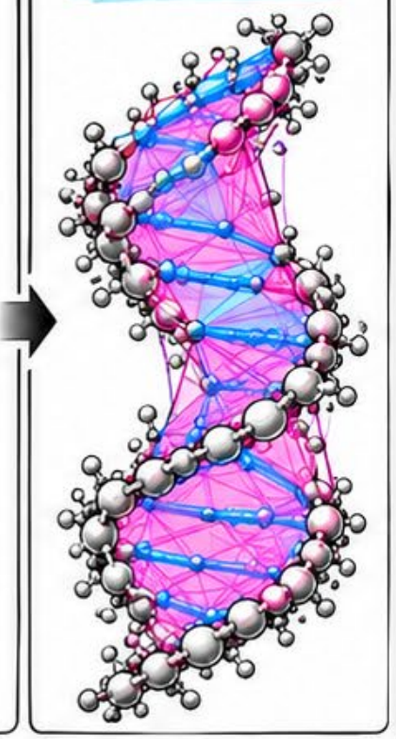
$\varepsilon = 3.5$ Å



$\varepsilon = 4.5$ Å



$\varepsilon = 6.0$ Å



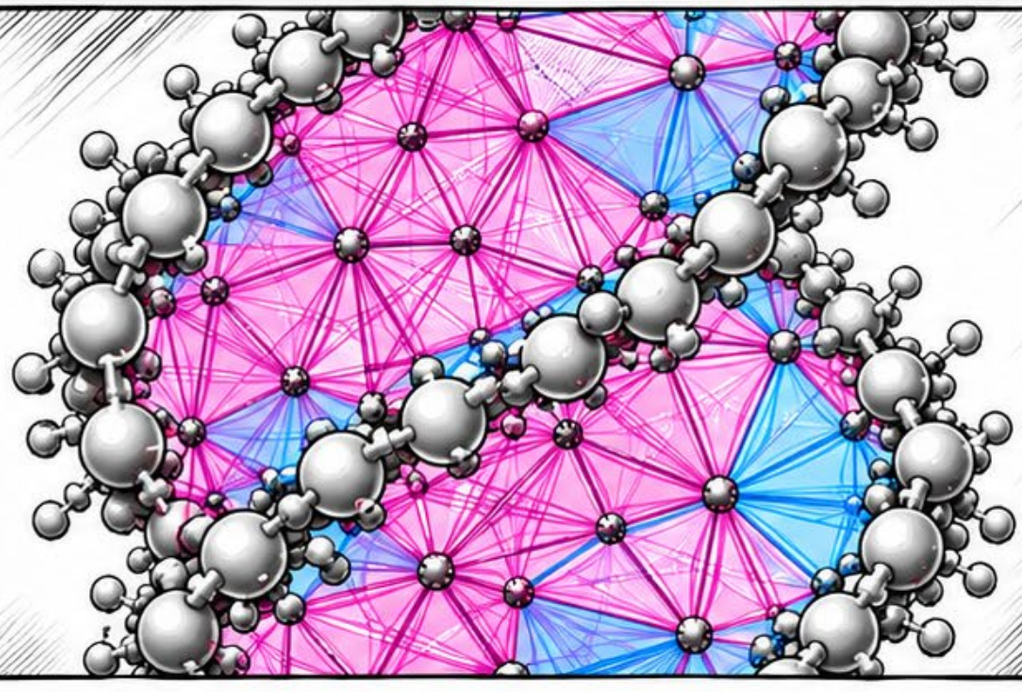
Limitation:

Dense complete graph

Computational complexity:

$$O(n2^n)$$

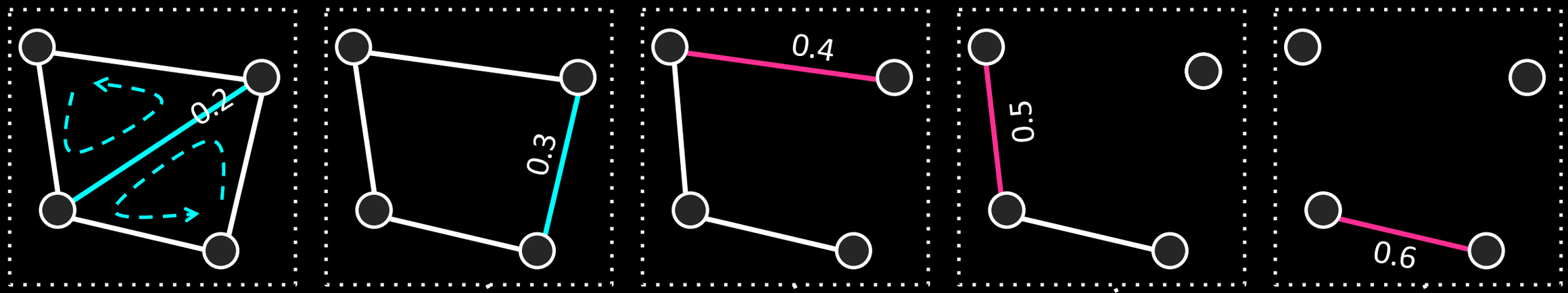
n : number of points (atoms)



Graph filtration (edge deletion from smallest to largest)

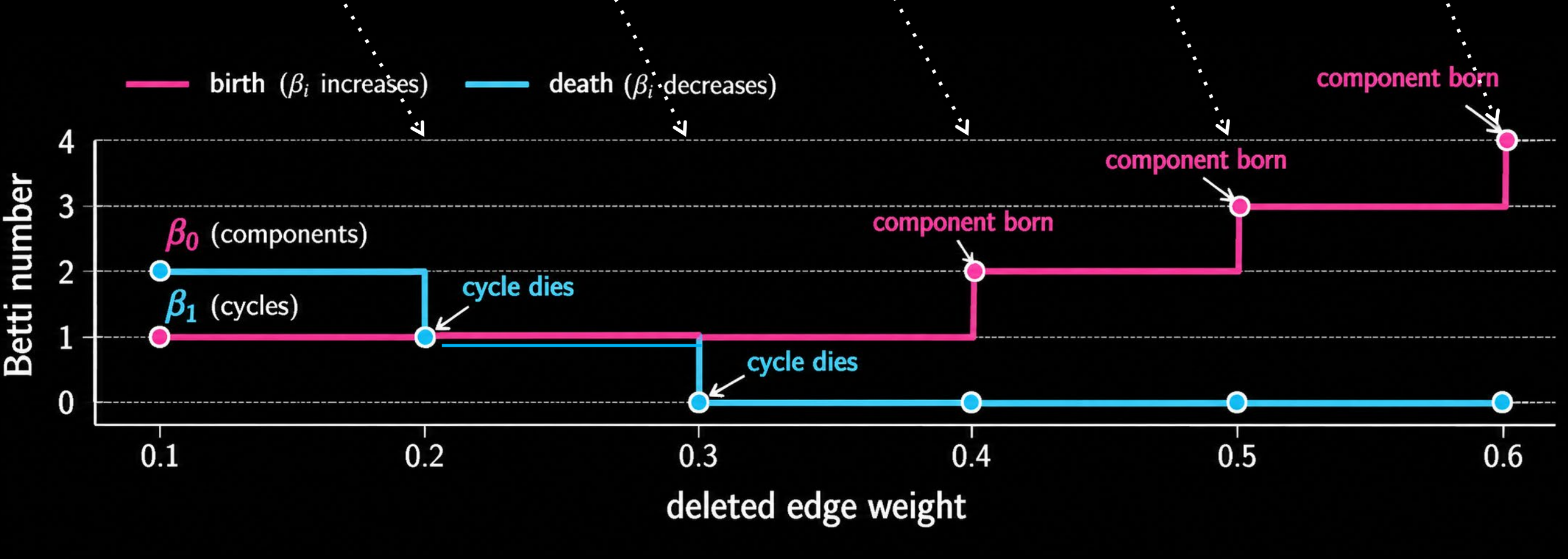
Deleting edges destroy cycles

Deleting edges create components



birth (β_i increases)

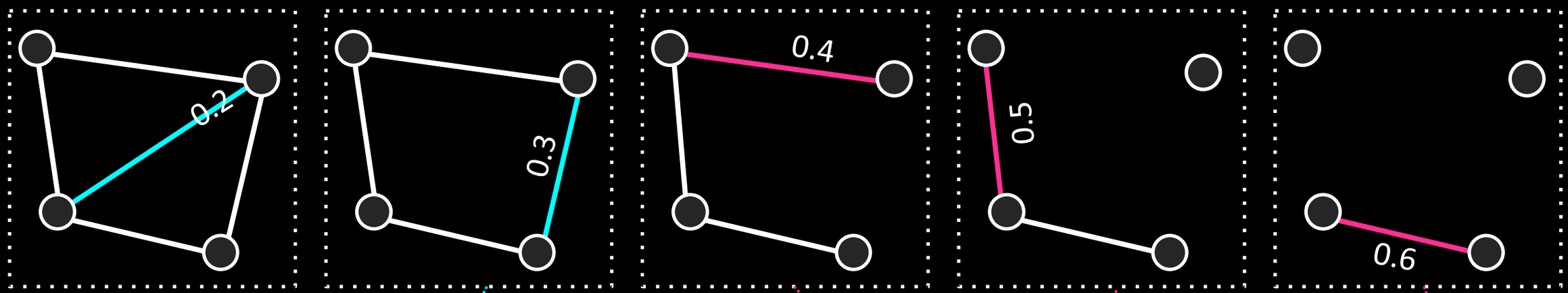
death (β_i decreases)

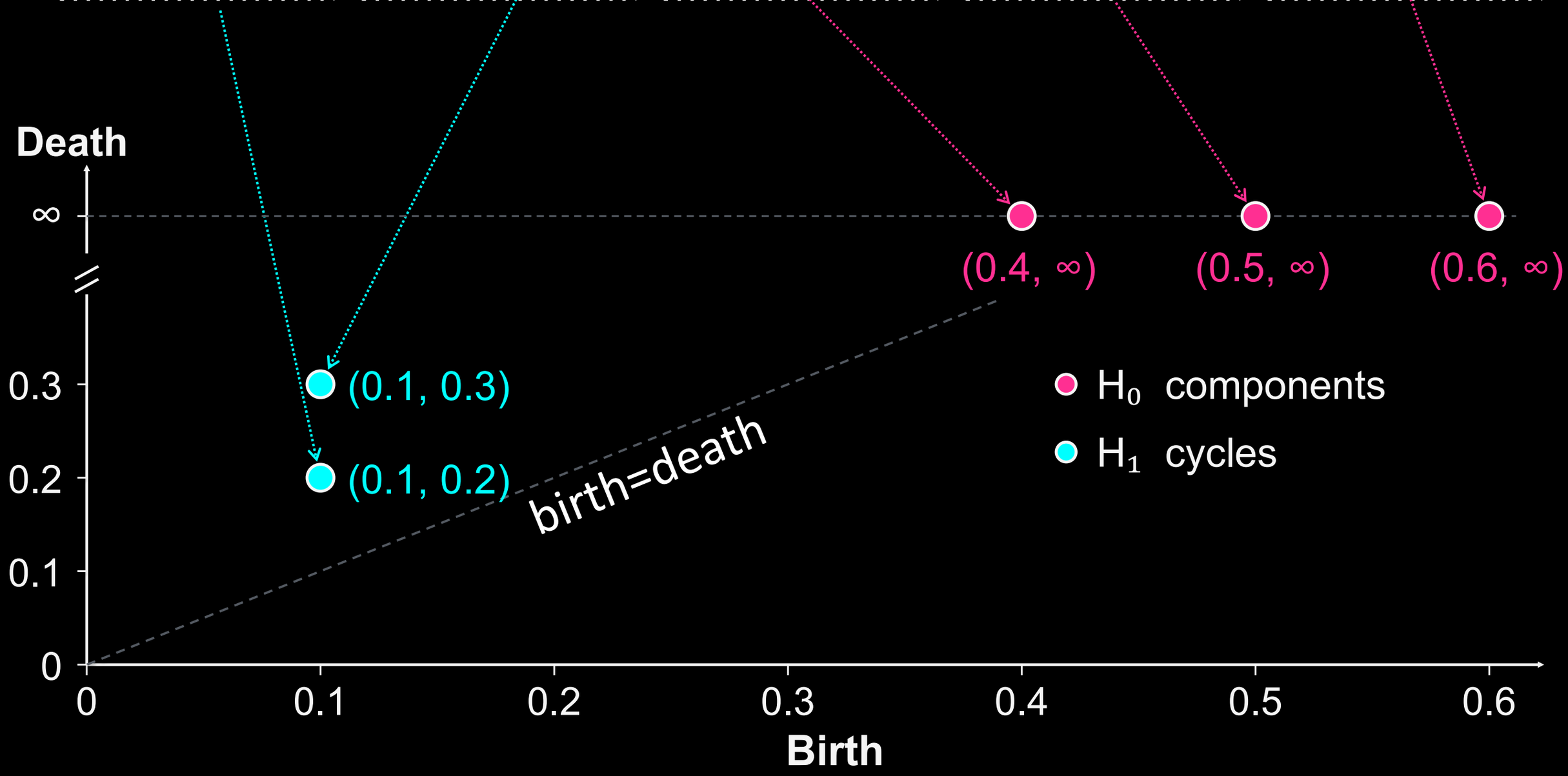


Persistent diagram (inefficient representation)

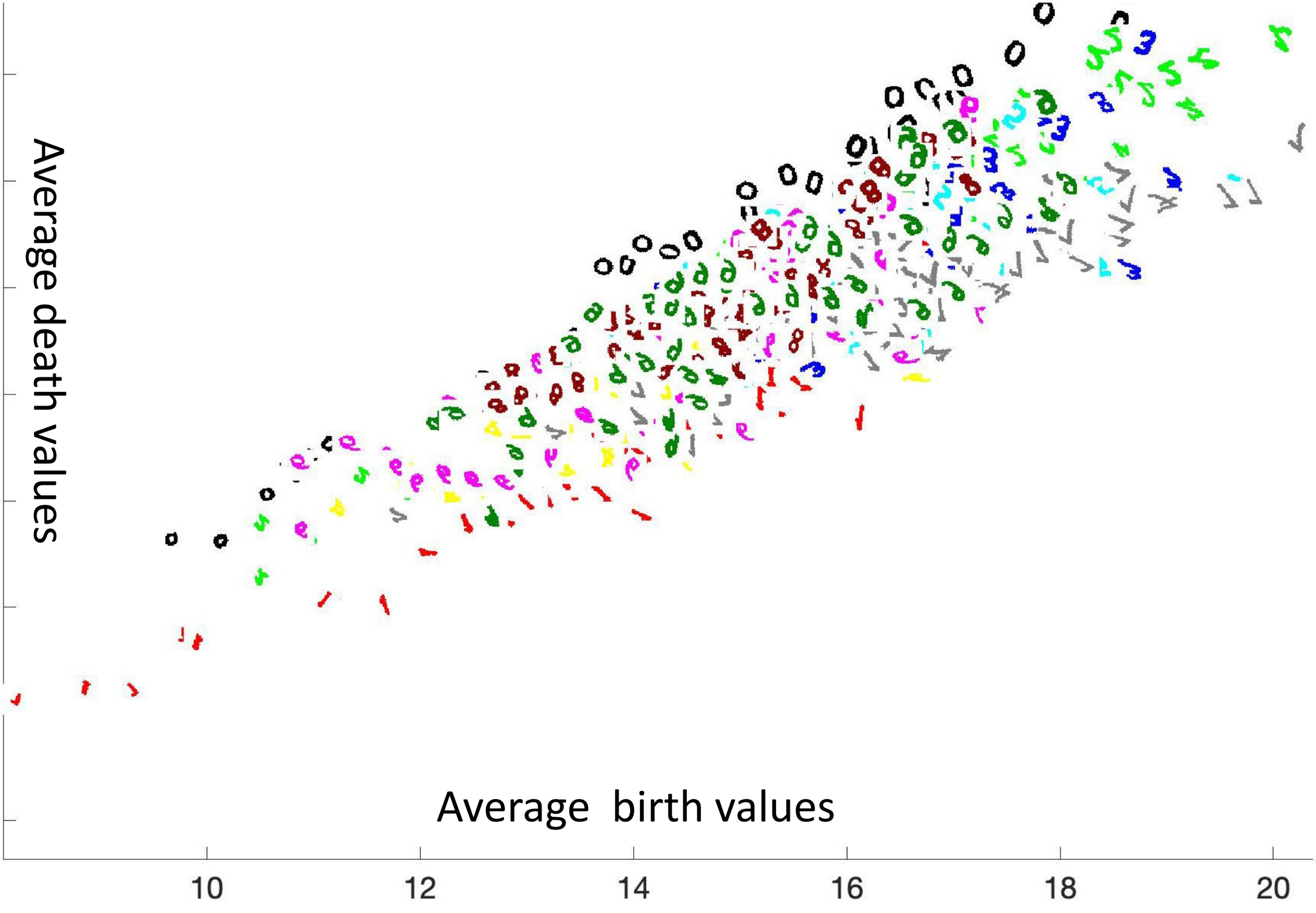
Deleting edges destroy cycles

Deleting edges create components



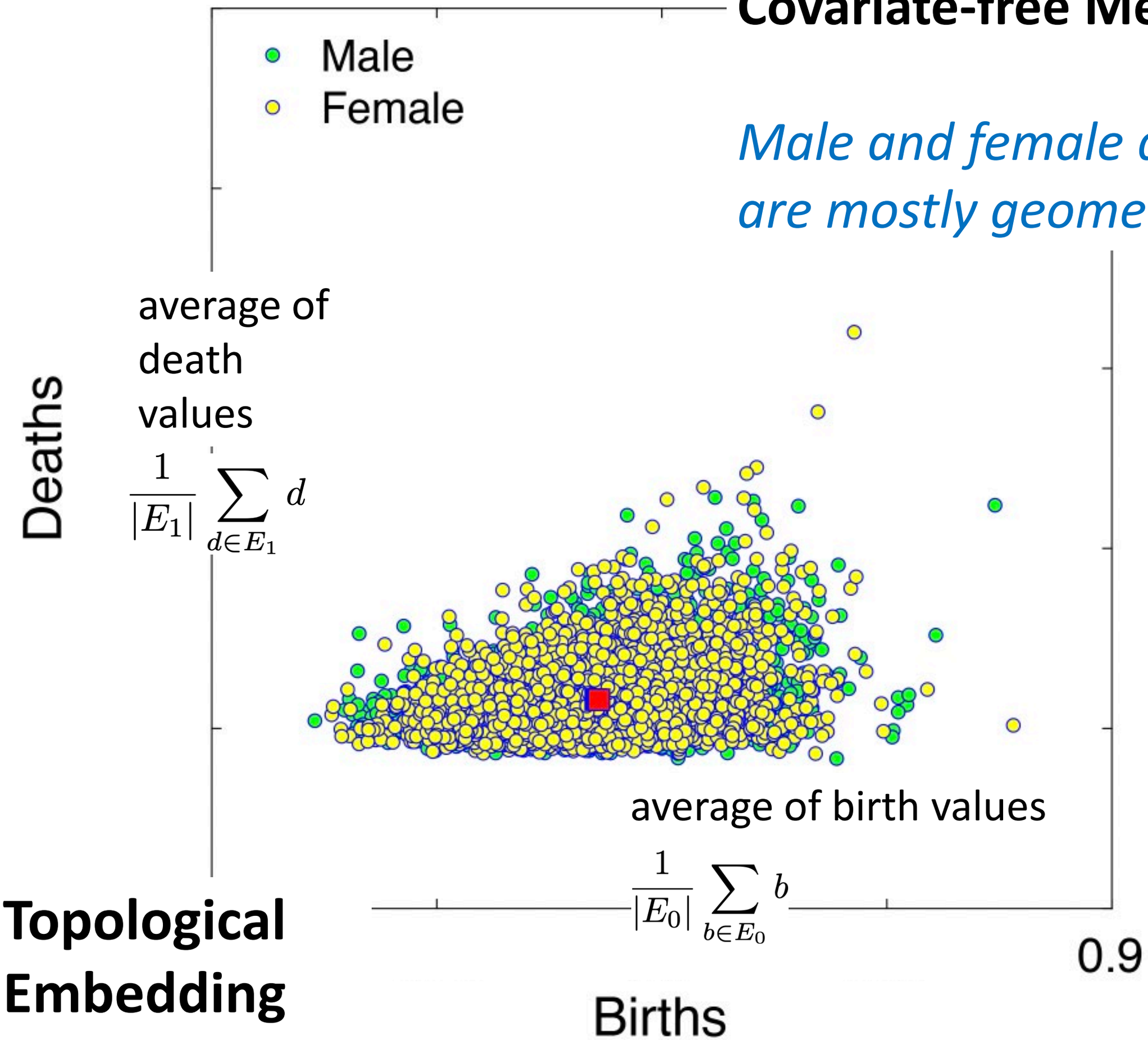


Topological Embedding on MINST



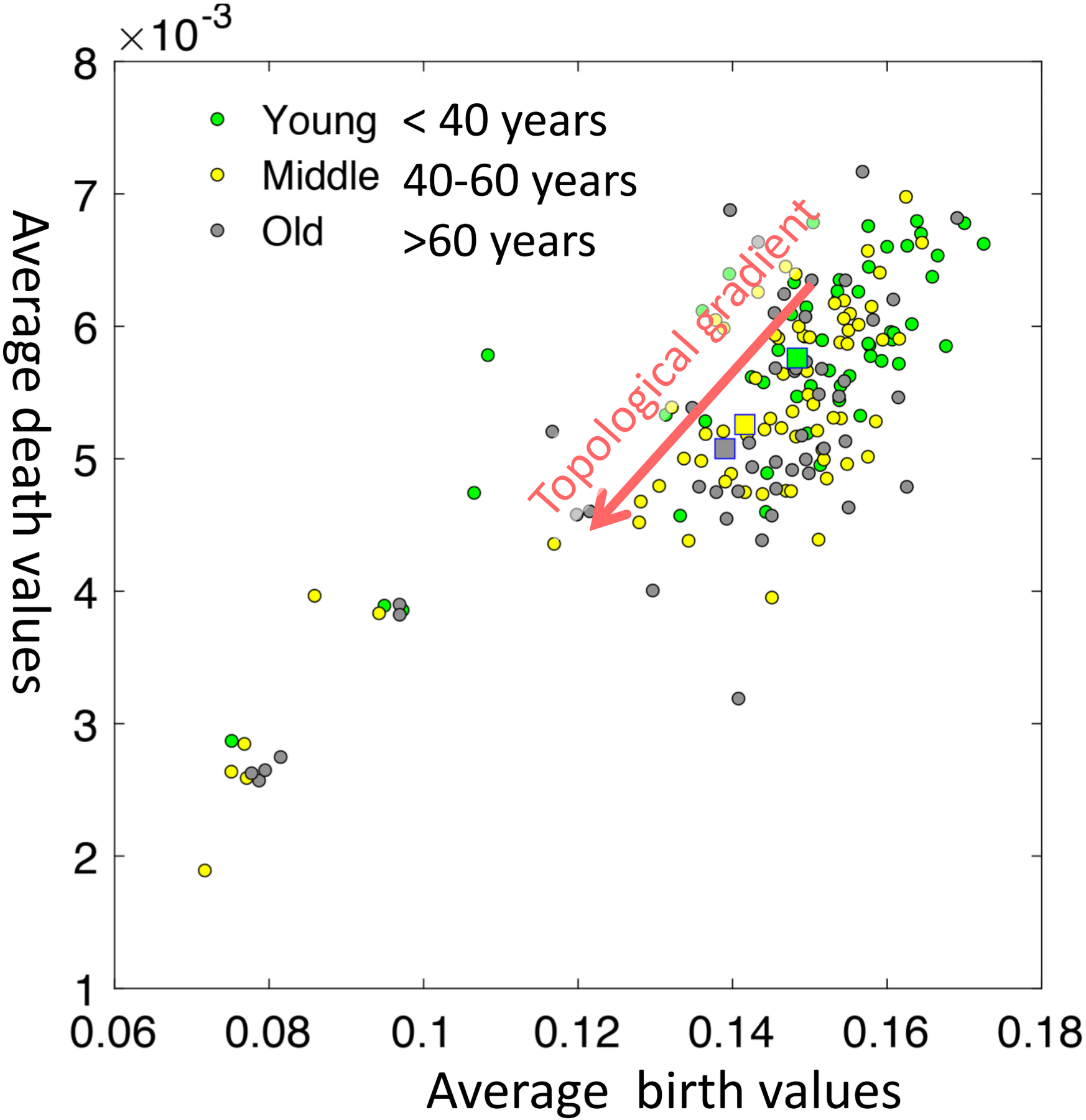
Covariate-free Method

Male and female differences are mostly geometric.



Topological Embedding

Topological Embedding *of Structural Connectivity (dMRI)*

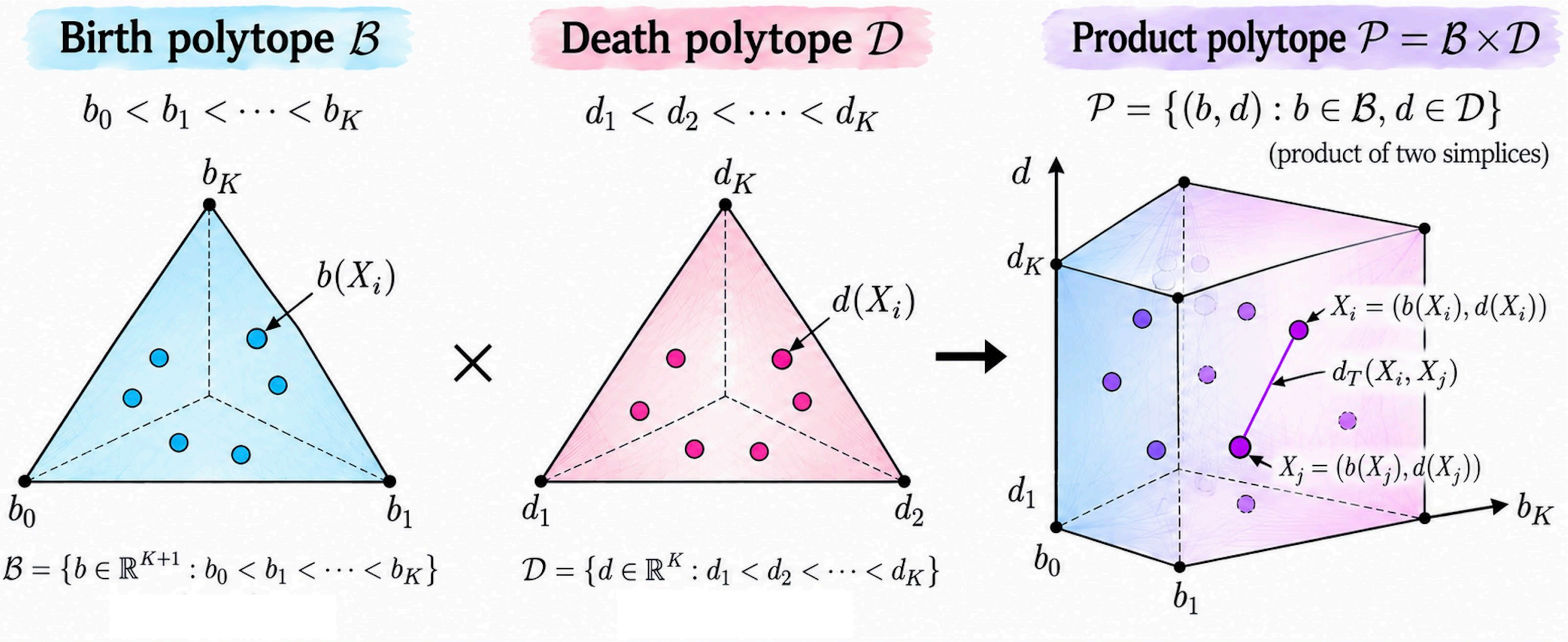


Data from

Jeffrey Phillips
James Gee
Univ. of
Pennsylvania

Aging is topological.

Topological distance → Euclidean distance

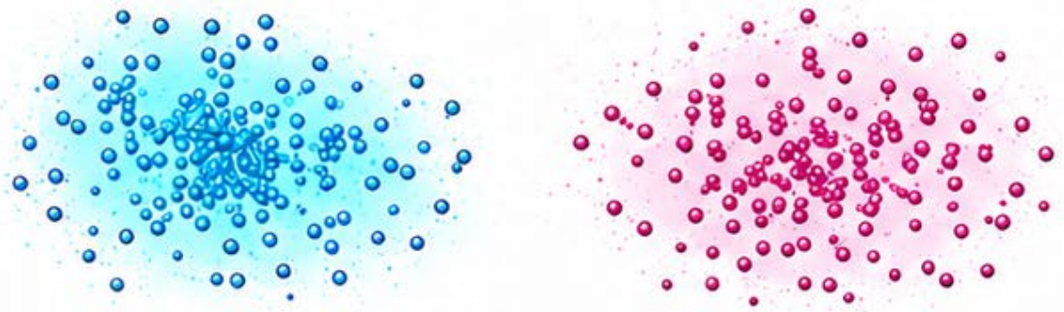


Run time for q number of edges

Chung et al. 2024 PLOS Computational

Wasserstein distance

$X \sim f_1$ $Y \sim f_2$



$P_1 = \{x_1, \ldots, x_q\} \subset \mathbb{R}^2$ $P_2 = \{y_1, \ldots, y_q\} \subset \mathbb{R}^2$

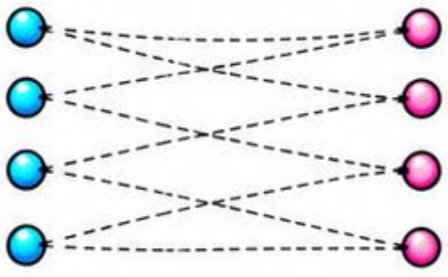
$$f_1(x) = \frac{1}{q}\sum_{i=1}^{q}\delta(x - x_i) \qquad f_2(y) = \frac{1}{q}\sum_{i=1}^{q}\delta(y - y_i)$$

r-Wasserstein distance:

$$d(\mu,\nu)^r = \inf_{(X,Y)} \mathbb{E}\,\|X - Y\|^r$$

2-Wasserstein distance:

$$d(P_1, P_2) = \inf_{\psi: P_1 \to P_2} \left(\sum_{x \in P_1} \|x - \psi(x)\|^2\right)^{1/2}$$

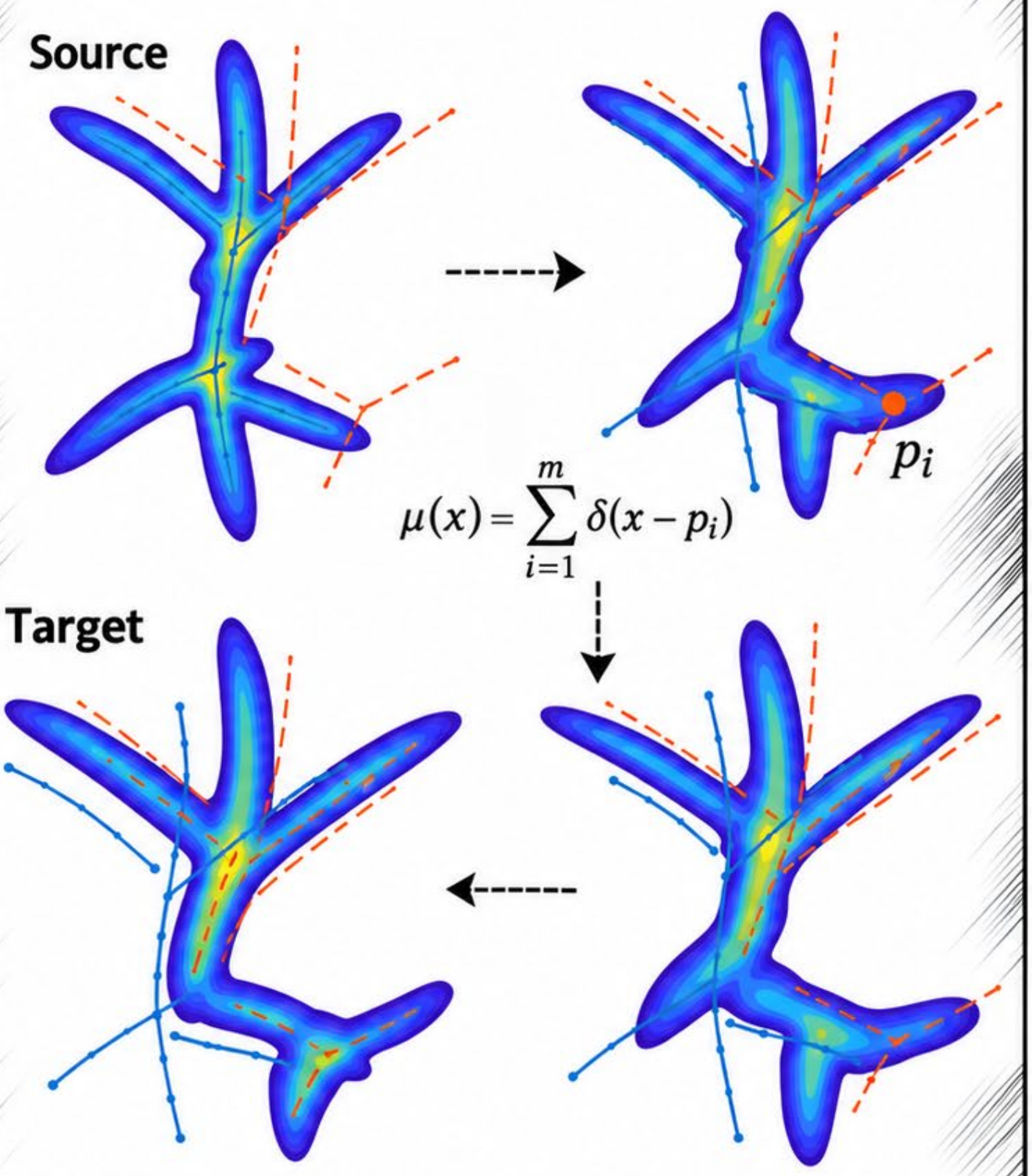


Optimal bijection ψ^*

Assignment problem:
Hungarian algorithm

$\mathcal{O}(q^3)$

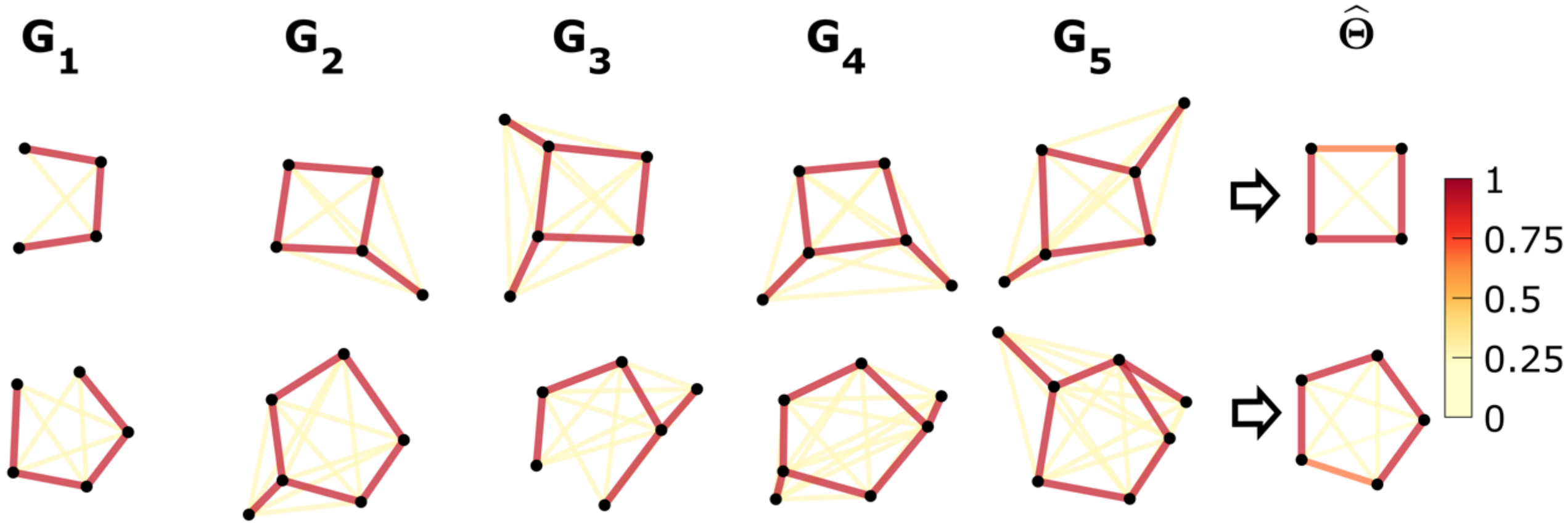
Wasserstein registration of trees



Chung 2026, arXiv:2608.04770

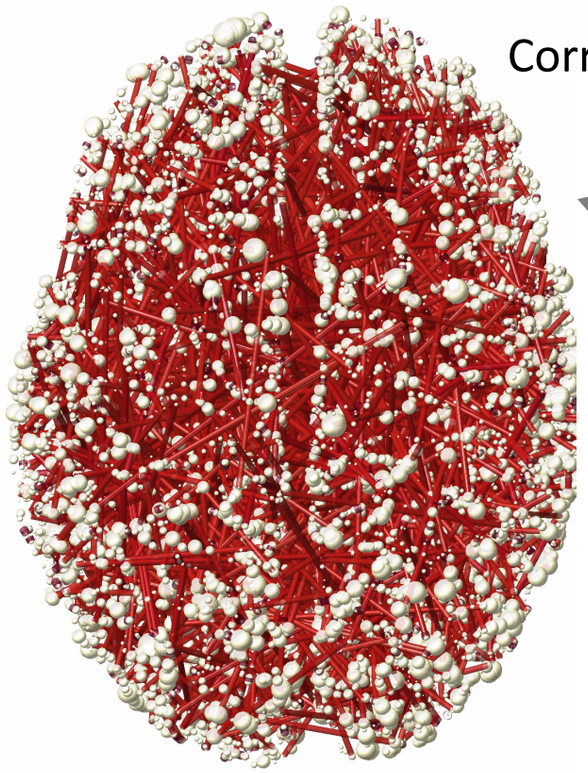
Topological averaging

= minimizer of topological distance

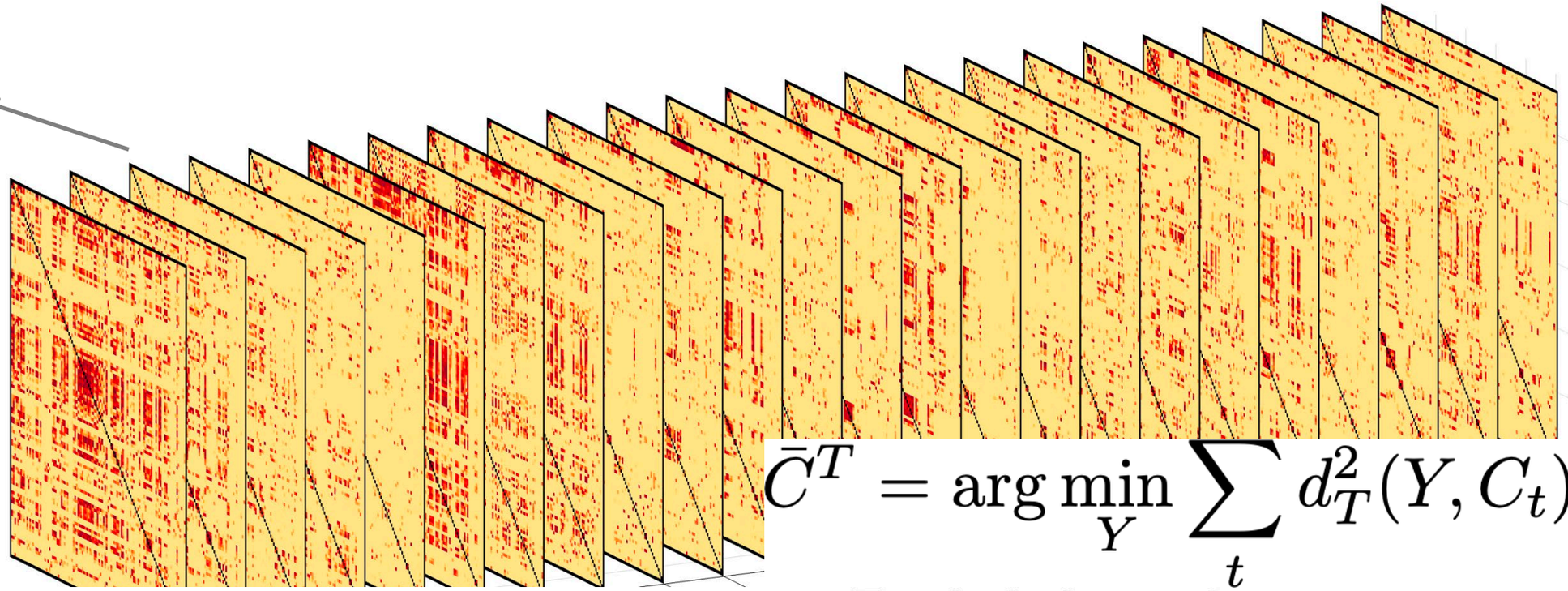


Songdechakraiwut and Chung 2023 Annals of Applied

We can average rs-fMRI across time and subjects (n=400,15mn)

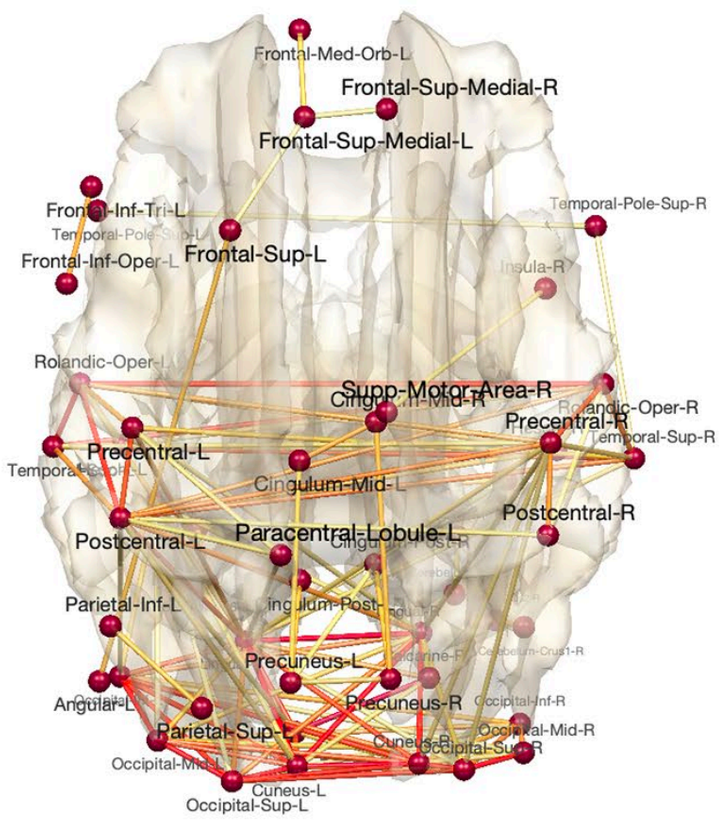


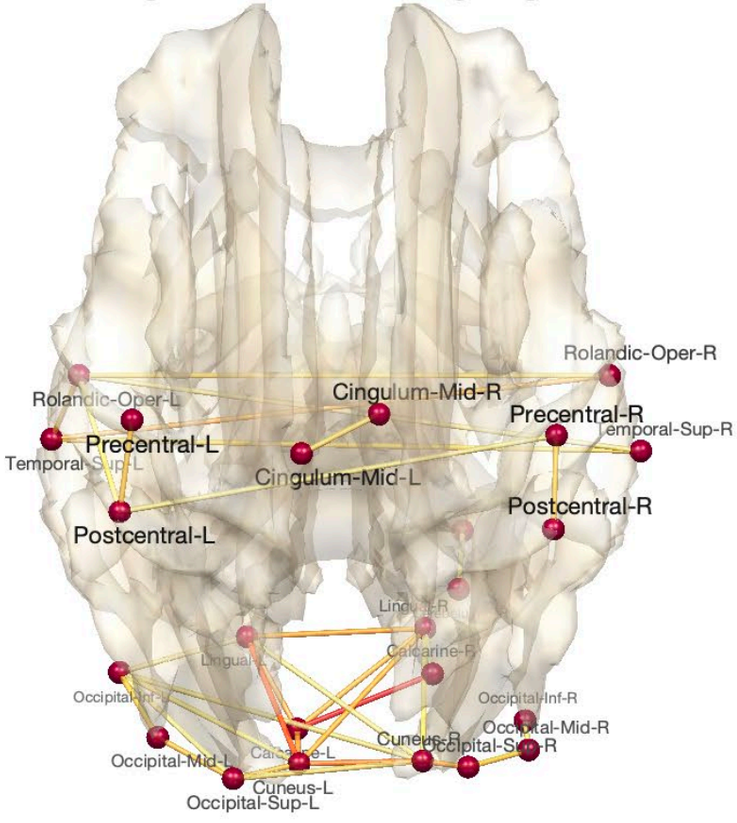
Correlation above 0.7

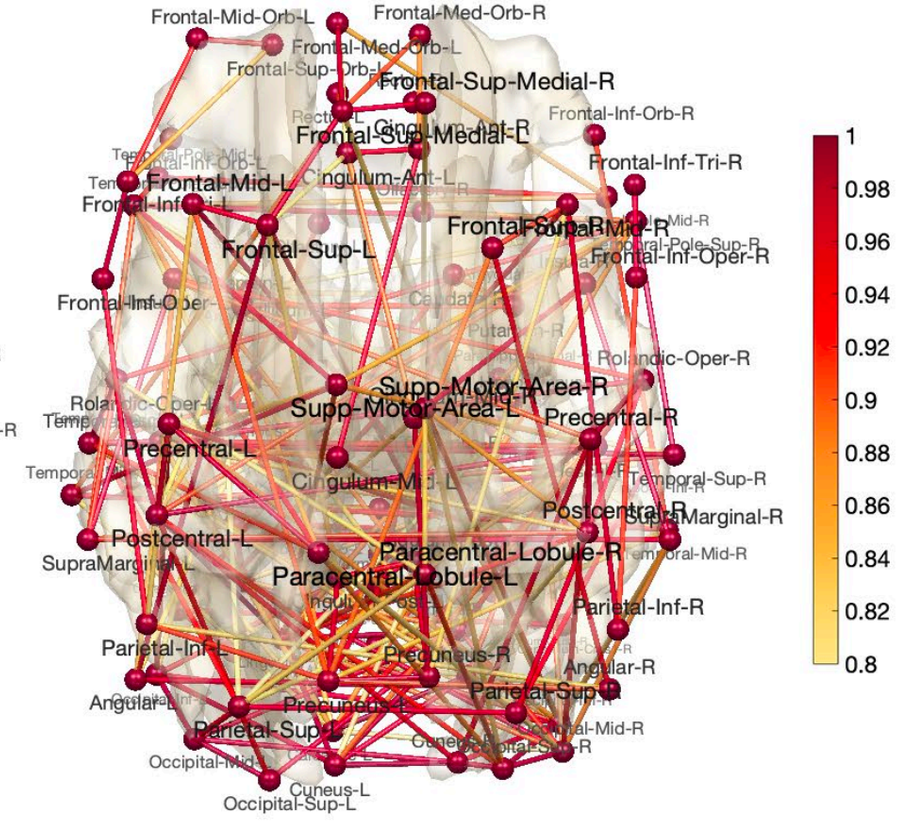


$$\bar{C}^T = \arg\min_Y \sum_t d_T^2(Y, C_t)$$

Correlation **Algebraic avearging** **Topological averaging**

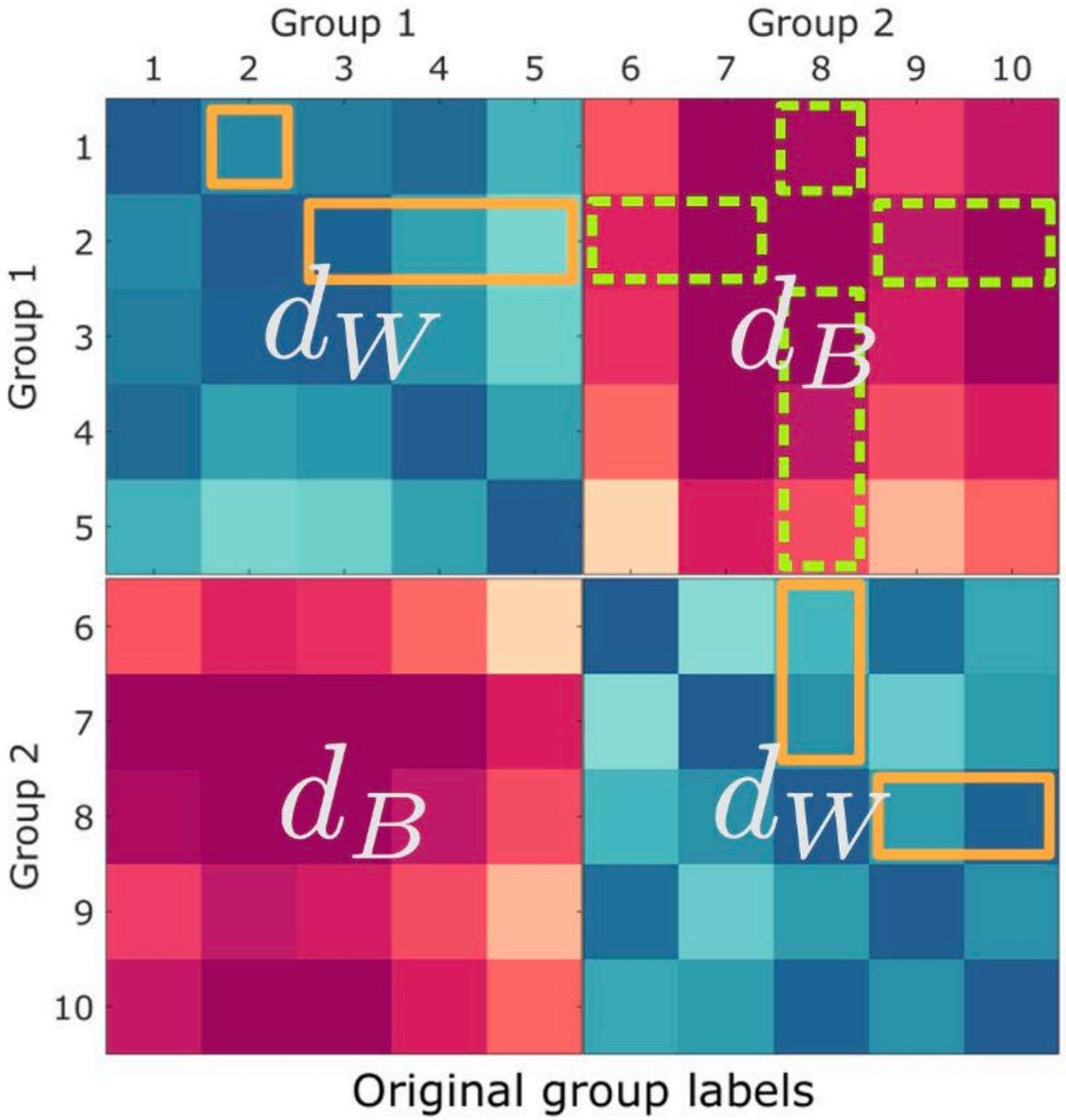




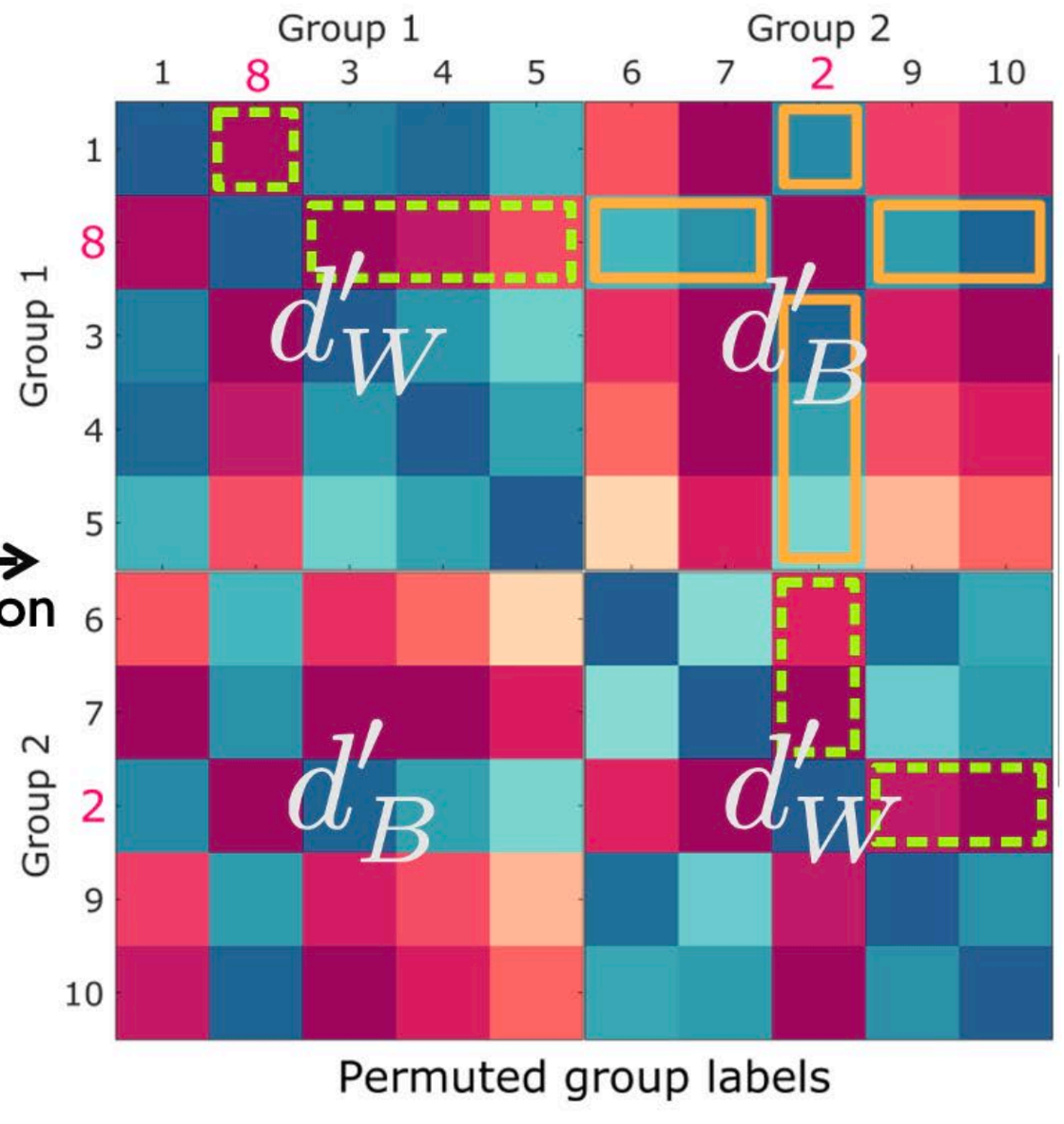


Sample variance 0.47+-0.02 0.02+-0.01

Topological inference via online permutation test



Transposition



Distance matrix precomputed.
A permutation only rearrange terms

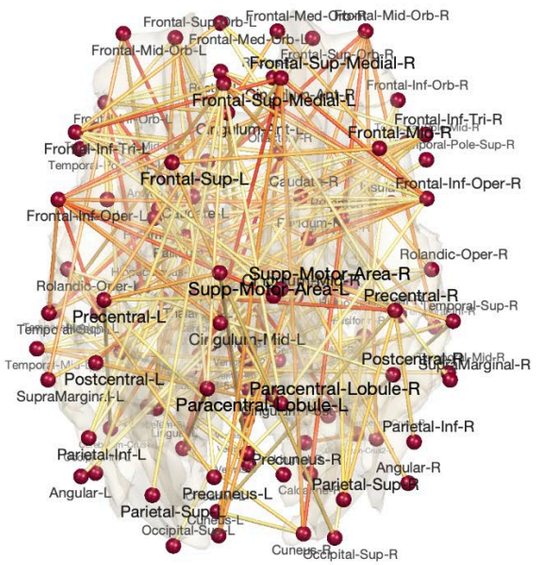
100 million
transpositions < 2 mins

Songdechakraiwut & Chung 2023 Annals of Applied Stat.

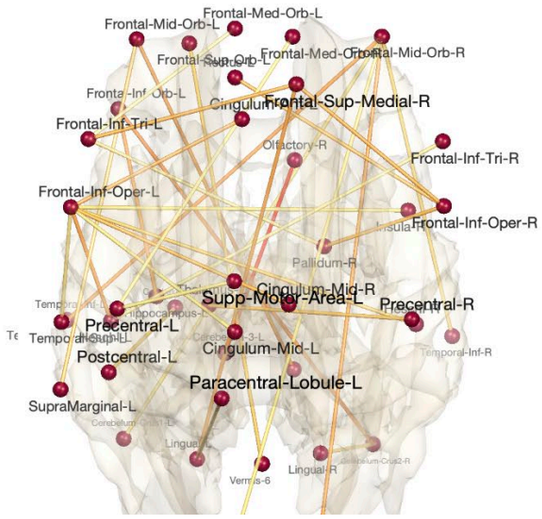
Topological clustering

rs-fMRI connectivity (n=479, UW-Madison)

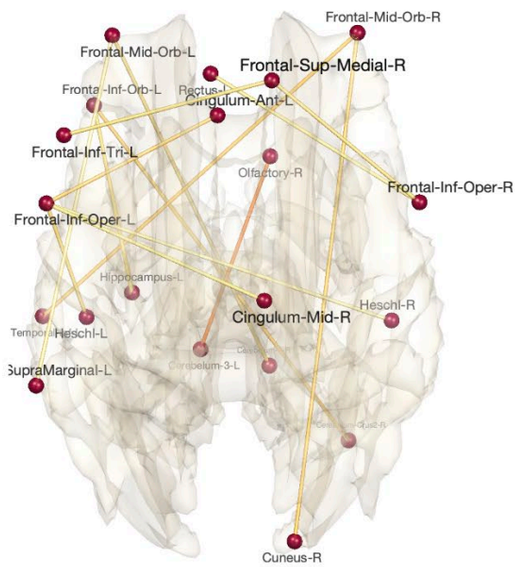
State 1



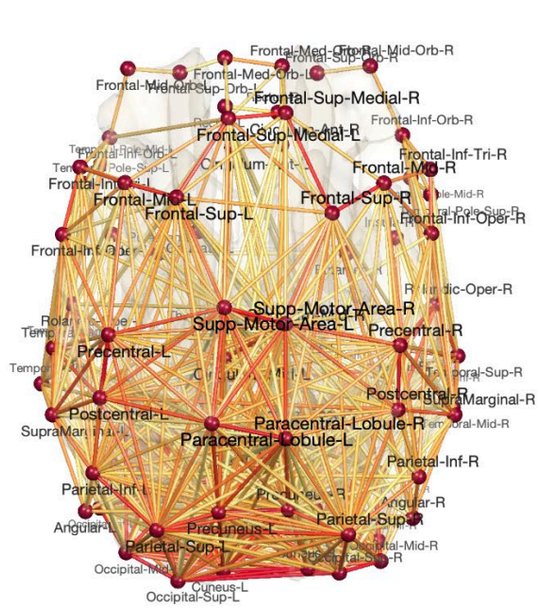
State 2

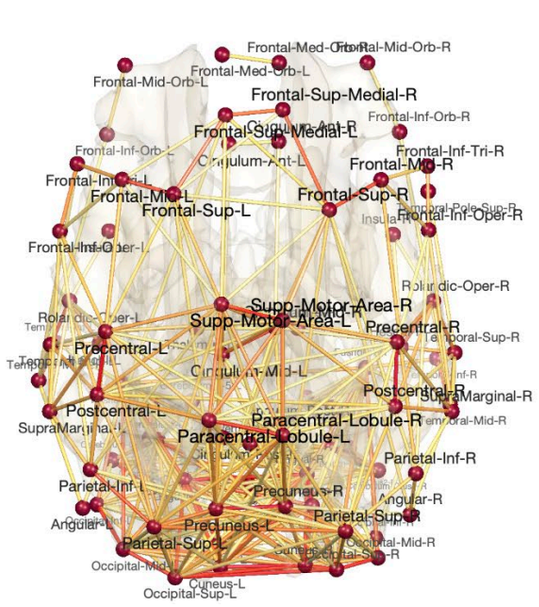


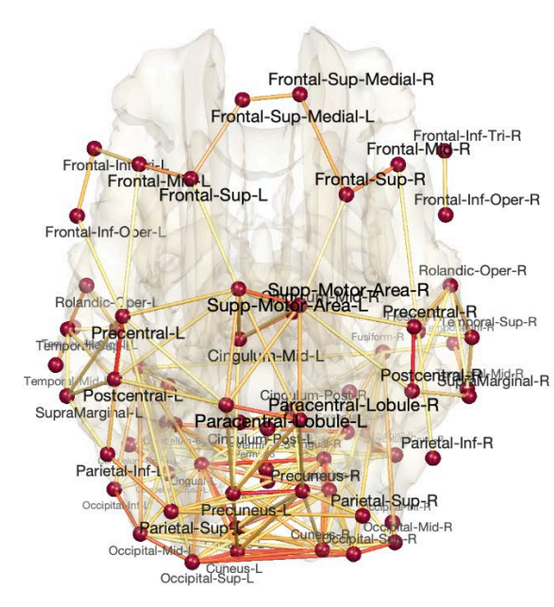
State 3



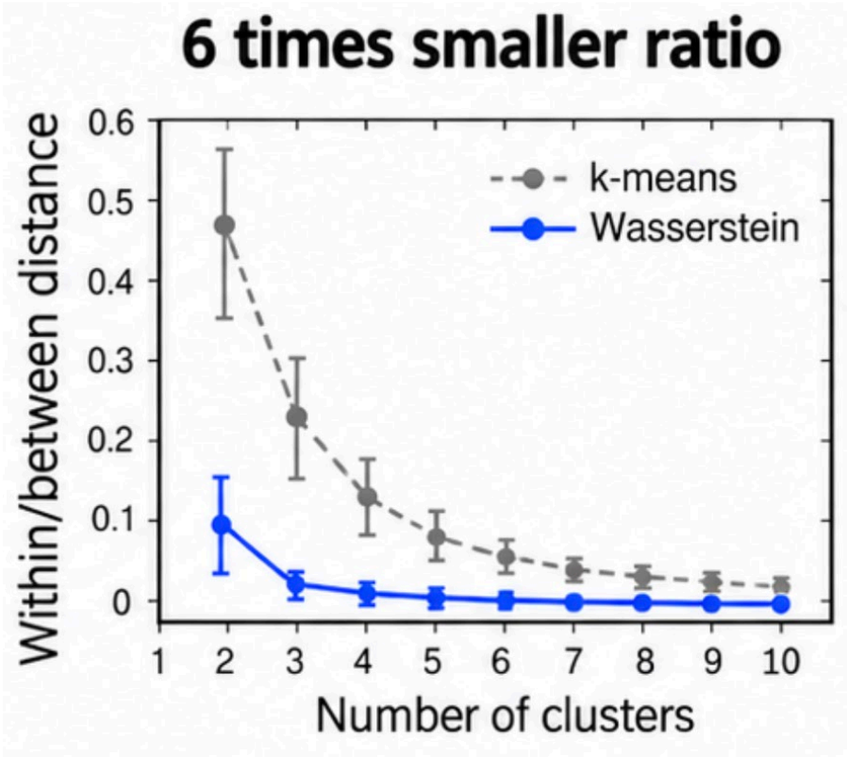
k-means







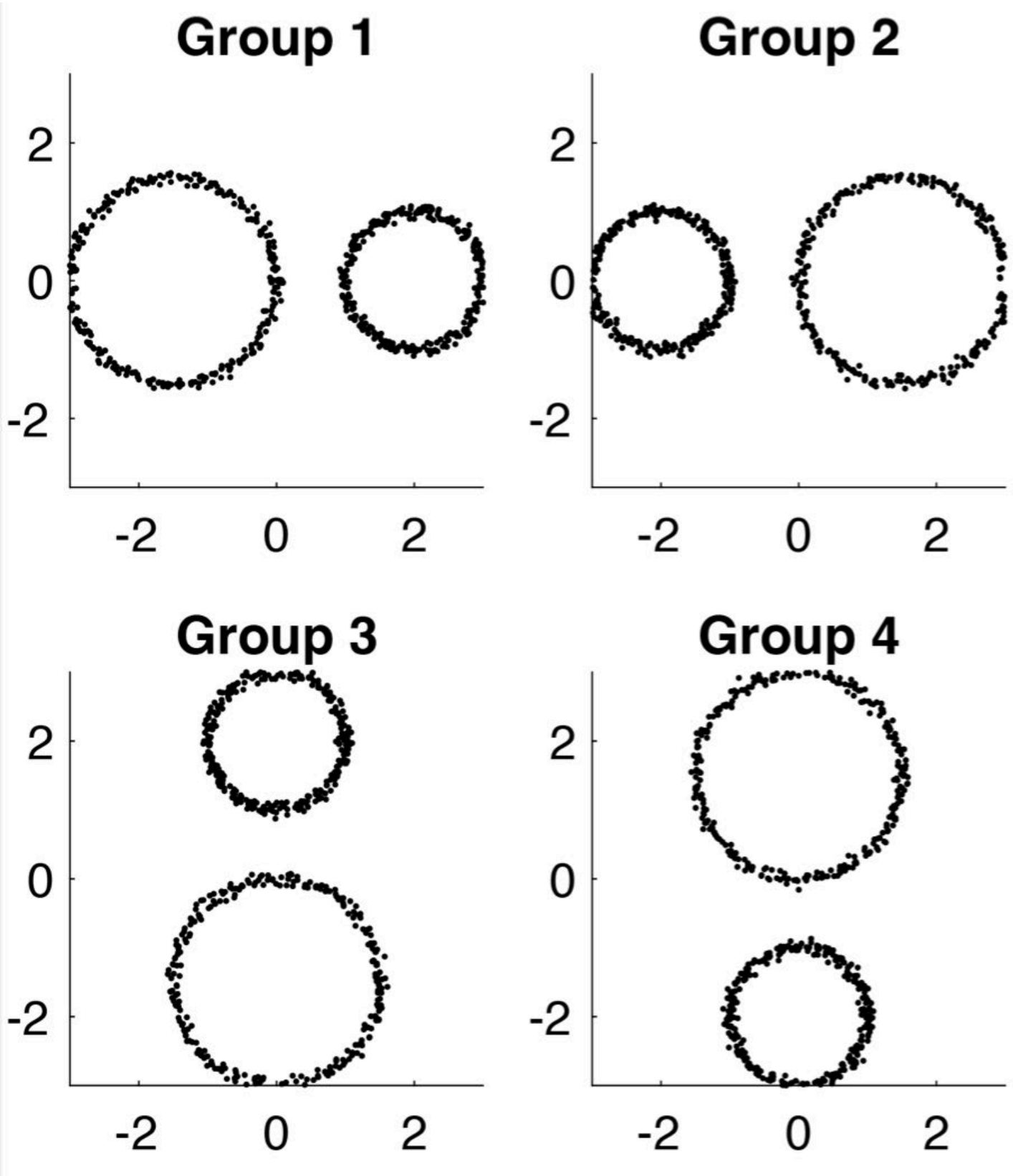




Wasserstein

Chung at al. 2024 PLOS Computational Blology

Existing clustering fails in topological task



K-means clustering 0.98 +- 0.01

Hierarchical clustering 1.00 +- 0.00

Topological clustering 0.63 +- 0.04

Spectral TDA workflow

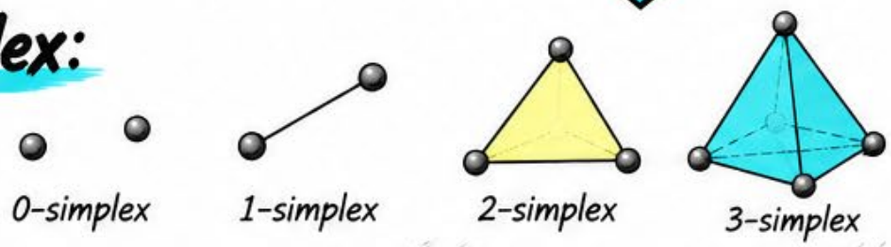
Data:

functional, point cloud, images, graphs

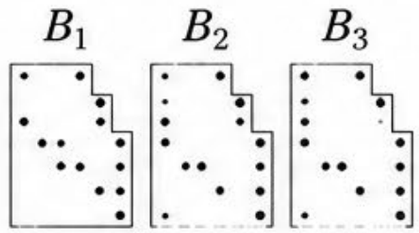


Simplicial complex:

boundary matrices

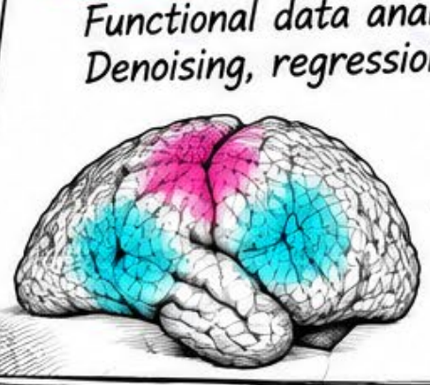


$B_k : C_k \to C_{k-1}$

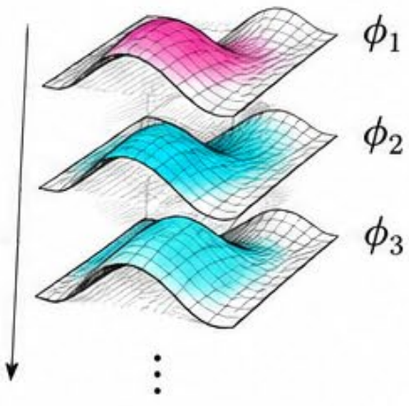


Hodge Expansion

Functional data analysis
Denoising, regression



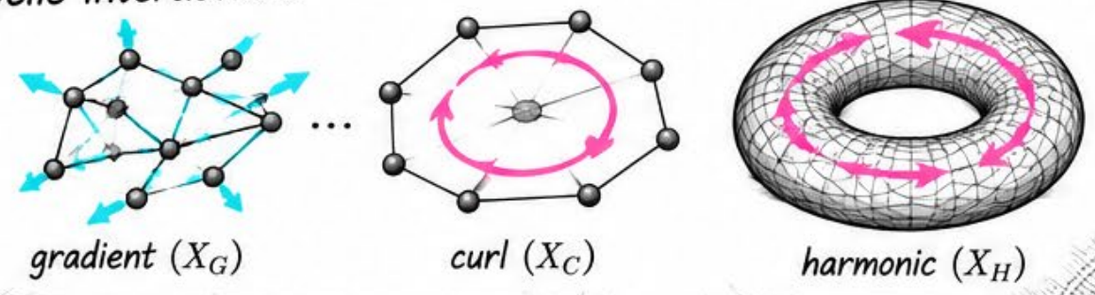
$$f = \sum_i \alpha_i \phi_i$$



Hodge decomposition

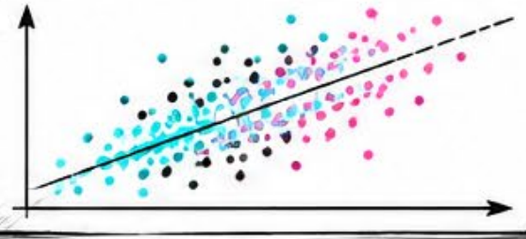
Directed interactions,
cyclic interactions

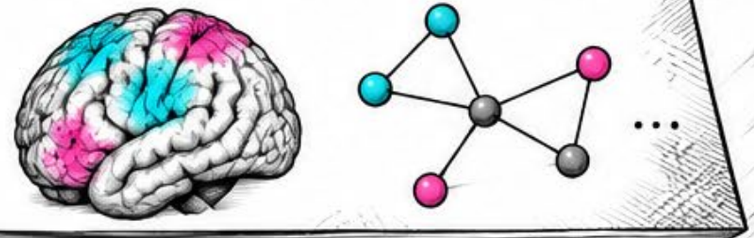
$$X = X_G + X_C + X_H$$



Inference & learning

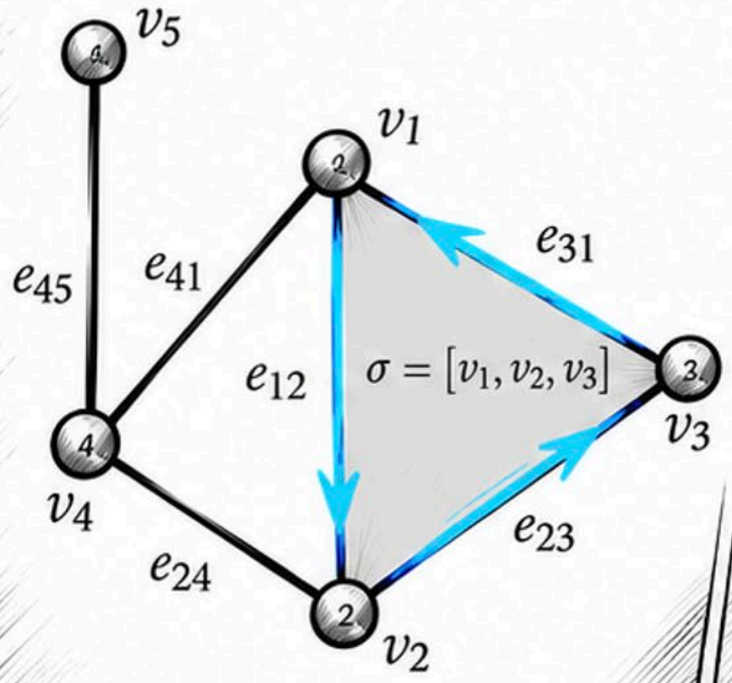
Statistical inference
Prediction, classification





Chung, El-Yaagoubi, Omao, 2026 ***Vector Space of Cycles****, arXiv:2606.08202*

Simplicial complex



Boundary matrix B_2

(2-simplices → 1-simplices)

$$\begin{array}{c} \\ e_{12} \\ e_{23} \\ e_{31} \\ e_{24} \\ e_{41} \\ e_{45} \end{array} \begin{array}{c} \sigma \\ \left[\begin{array}{c} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{array}\right. \end{array} \quad \partial\sigma = e_{12} + e_{23} + e_{31}$$

Coboundary matrix B_2^T

(1-simplices → 2-simplices)

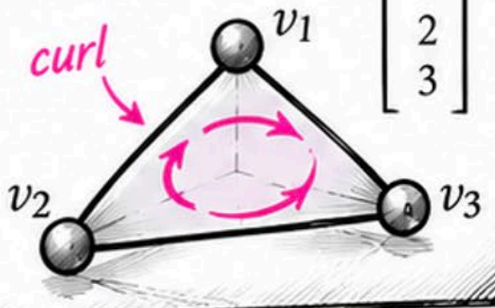
$$B_2^T = \begin{array}{c} \begin{array}{cccccc} e_{12} & e_{23} & e_{31} & e_{24} & e_{41} & e_{45} \end{array} \\ \left[\begin{array}{cccccc} 1 & 1 & 1 & 0 & 0 & 0 \end{array}\right] \end{array}$$

Edge signal $X \in \mathbb{R}^6$

$$\begin{array}{c} \\ e_{12} \\ e_{23} \\ e_{31} \\ e_{24} \\ e_{41} \\ e_{45} \end{array} \begin{array}{c} \\ \left[\begin{array}{c} 1 \\ 2 \\ 5 \\ 1 \\ 2 \\ 3 \end{array}\right. \end{array}$$

Face value *(sum over edges)*

$$B_2^T X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 5 \\ 1 \\ 2 \\ 3 \end{bmatrix} = 8$$



Generalization to B_k

$$B_k \qquad B_k^T$$

$$\cdots \begin{bmatrix} \pm 1 & 0 & \cdots & 0 \\ 0 & \pm 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pm 1 \end{bmatrix} \qquad \begin{bmatrix} \pm 1 & 0 & \cdots & \\ 0 & \pm 1 & \cdots & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \pm 1 \end{bmatrix}$$

faces → $(k-1)$-faces (boundary matrix)

$(k-1)$-faces → faces (coboundary matrix)

Anand and Chung, 2022 IEEE

k-th Hodge Laplacian

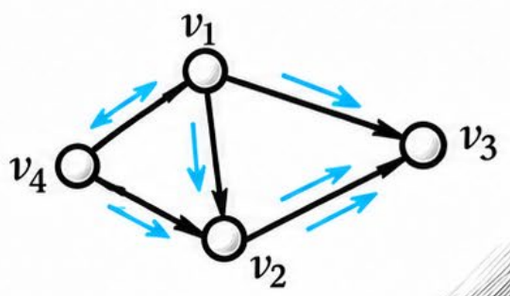
$$\Delta_1 = \mathrm{B}_1^\top \mathrm{B}_1 + \mathrm{B}_2 \mathrm{B}_2^\top$$

($\mathrm{B}_1^\top$: # of nodes; B_1: # of edges; B_2: # of edges; $\mathrm{B}_2^\top$: # of faces)

k = 0 : Graph Laplacian

$$\Delta_0 = B_1 B_1^{\mathrm{T}}$$

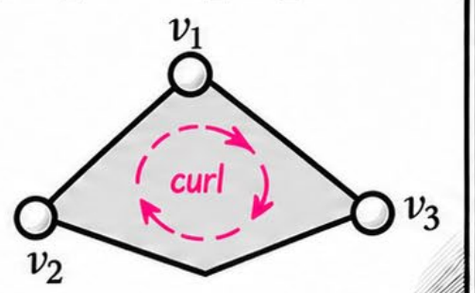
Acts on nodes (0-simplices). Measures net outflow at each node.



k = 1 : Edge Laplacian

$$\Delta_1 = B_1^{\mathrm{T}} B_1 + B_2 B_2^{\mathrm{T}}$$

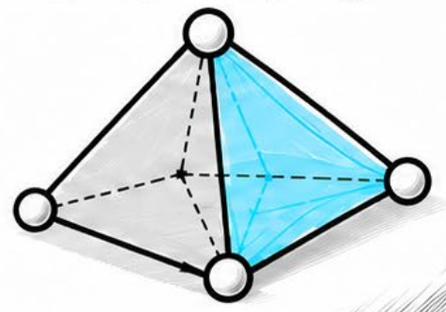
Acts on edges (1-simplices). Measures both divergence and curl of edge flows.



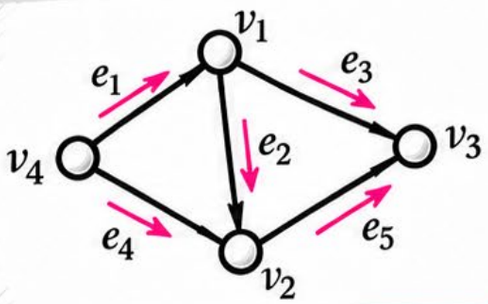
k = 2 : Face Laplacian

$$\Delta_2 = B_2^{\mathrm{T}} B_2 + B_3 B_3^{\mathrm{T}}$$

Acts on triangles (2-simplices). Measures how face values are inconsistent across tetrahedra (3-simplices).



Example (k = 1)



Edge $e_3 = (v_1 \to v_3)$ is included.

B_1 (edges × nodes)

	v_1	v_2	v_3	v_4
e_1	−1	0	0	1
e_2	1	−1	0	0
e_3	−1	0	1	0
e_4	0	−1	0	1
e_5	0	1	−1	0

Each column has -1 at the tail and $+1$ at the head of the edge.

B_2 (faces × edges)

	e_1	e_2	e_3	e_4	e_5
f_1	1	1	−1	0	0

Triangle $f_1 = [v_1, v_2, v_3]$ has oriented boundary $\partial f_1 = e_1 + e_2 - e_3$.

Edge Laplacian matrix

$$\Delta_1 = B_1^{\mathrm{T}} B_1 + B_2 B_2^{\mathrm{T}}$$

	e_1	e_2	e_3	e_4	e_5
e_1	2	1	−1	0	0
e_2	1	2	−1	0	0
e_3	−1	−1	2	0	0
e_4	0	0	0	1	−1
e_5	0	0	0	−1	1

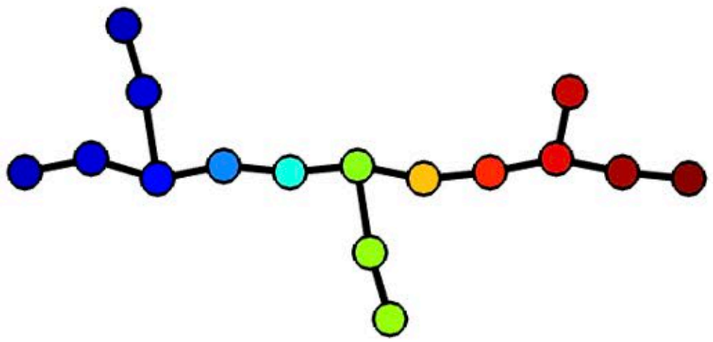
First 3 edges (e_1, e_2, e_3) form a cycle (curl), while e_4 and e_5 are tree edges. The Hodge Laplacian captures both divergence and curl structures.

Anand and Chung, 2023 IEEE

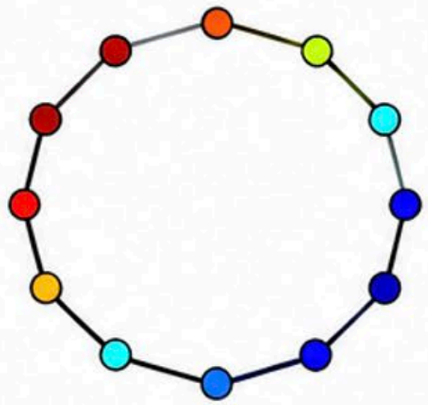
Eigenmodes of graph Laplacian vs. 1-Hodge Laplacian

Graph Laplacian (nodes)

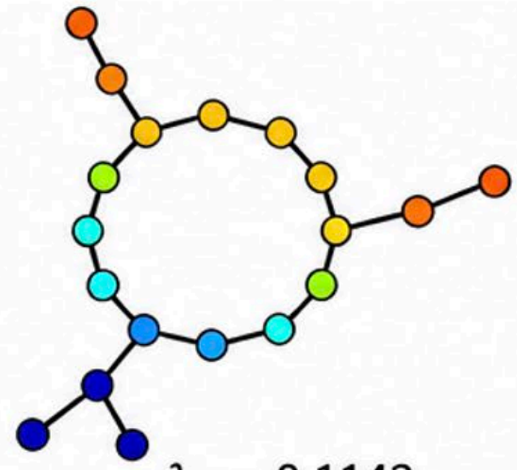
First nonzero eigenmode (λ_1)



$\lambda_1 = 0.0594$



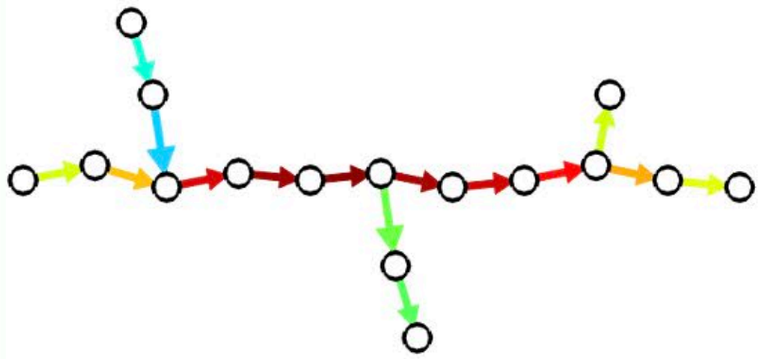
$\lambda_1 = 0.2679$



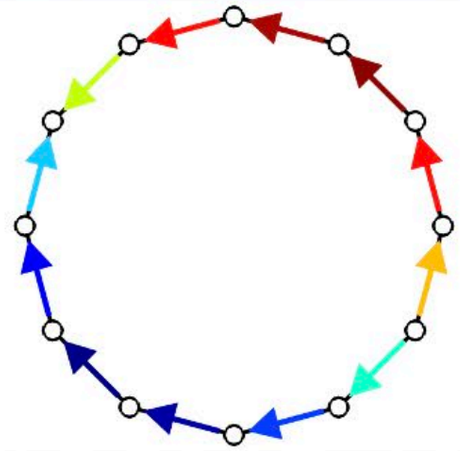
$\lambda_1 = 0.1142$

1-Hodge Laplacian (edges)

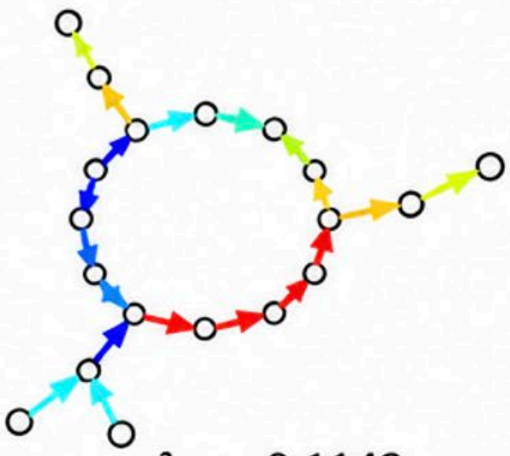
Gradient mode (λ_1)



$\lambda_1 = 0.0594$



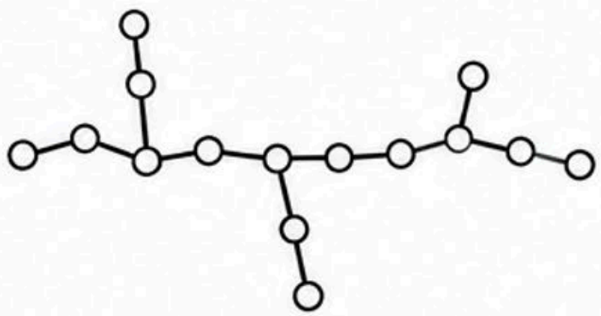
$\lambda_1 = 0.2679$



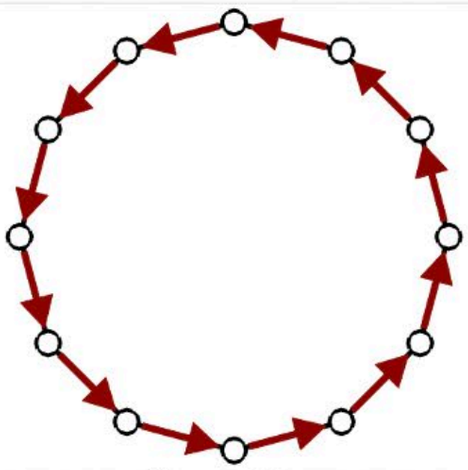
$\lambda_1 = 0.1142$

1-Hodge Laplacian (edges)

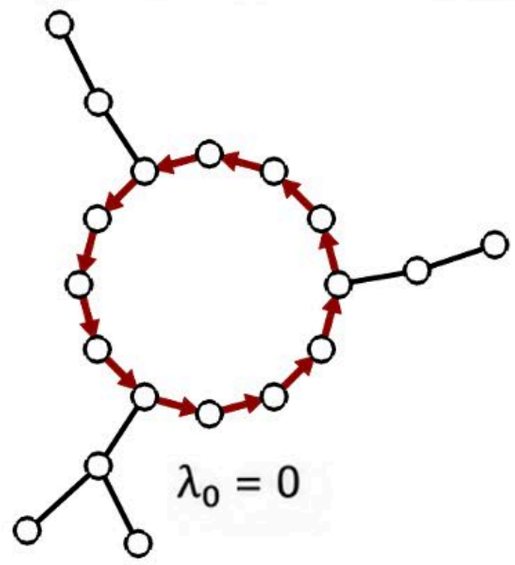
Harmonic mode ($\lambda_0 = 0$)



$\lambda_0 = 0$
(none)



$\lambda_0 = 0$



$\lambda_0 = 0$

Hodge decomposition

Edge flow X

$$X = X_G + X_C + X_H$$

Original edge flow X

Gradient (exact) *(flow from nodes)*

$$X_G = \mathrm{B}_1^T s$$

Function on nodes (scalar potential)

Curl (coexact) *(flow around triangles)*

$$X_C = \mathrm{B}_2 \phi$$

Function on triangles (triangle potential)

Harmonic *(topological cycles)*

$$X_H$$

Global, nontrivial cycles (in $\ker(\mathrm{B}_1) \cap \ker(\mathrm{B}_2^T)$)

Original X



Gradient X_G



Curl X_C



Harmonic X_H



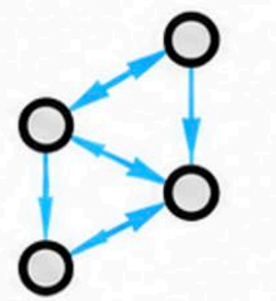
Least Squares Estimation of X_H

$$X_H = \arg \min_{Z \in \mathbb{R}^{|E|}} \|X - Z\|^2$$

subject to $\mathrm{B}_1 Z = 0, \quad \mathrm{B}_2^T Z = 0$

$\mathrm{B}_1 Z = 0$ *(no divergence at nodes)*

$\mathrm{B}_2^T Z = 0$ *(no curl on triangles)*



Zero net flow at each node

$$B = \begin{bmatrix} \mathrm{B}_1 \\ \mathrm{B}_2^T \end{bmatrix}$$



Zero circulation on each triangle

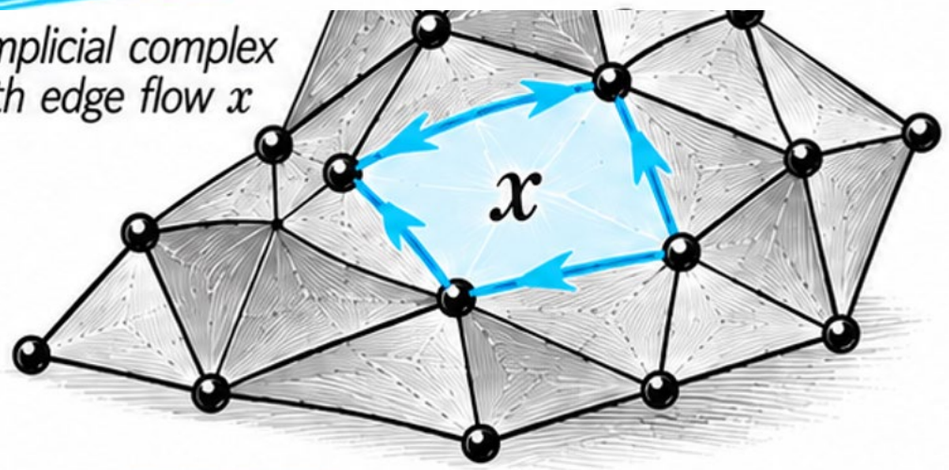
Orthogonal projection onto harmonic subspace

$$\mathcal{P}_H = I - B^T (BB^T)^+ B$$

$$X_H = \mathcal{P}_H X$$

El-Yaagoubi, Chung, Ombao

From boundary matrix to cycles and harmonics

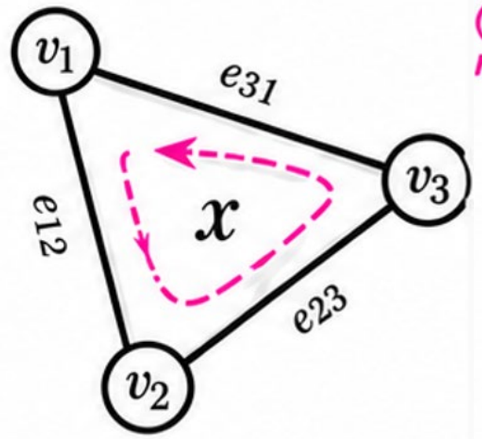


$$C_k = \ker(B_k) = \left\{ x \in \mathbb{R}^{|E_k|} : B_k x = 0 \right\}$$

k-cycles: edge flows with zero boundary (vertex divergence)

1-cycles (loops)

$$C_1 = \ker(B_1)$$

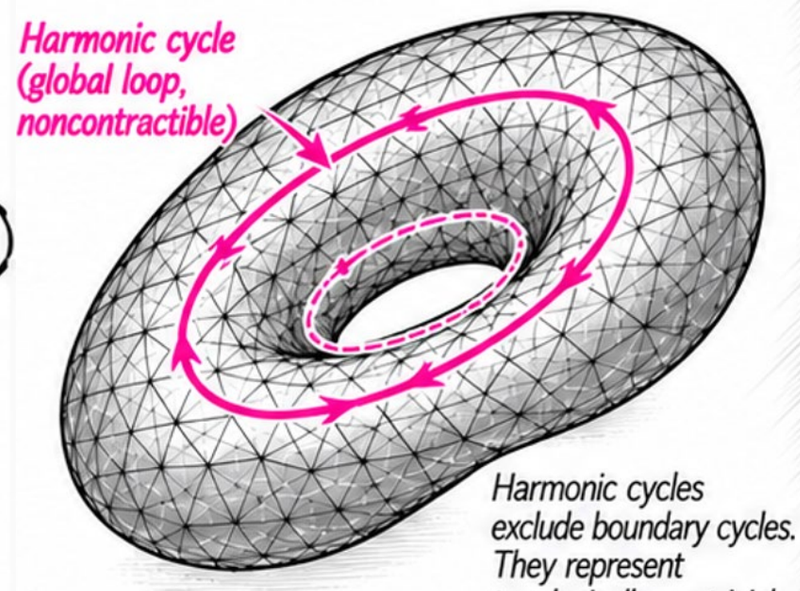


$$B_1 x = 0$$

(zero net flow at each node)

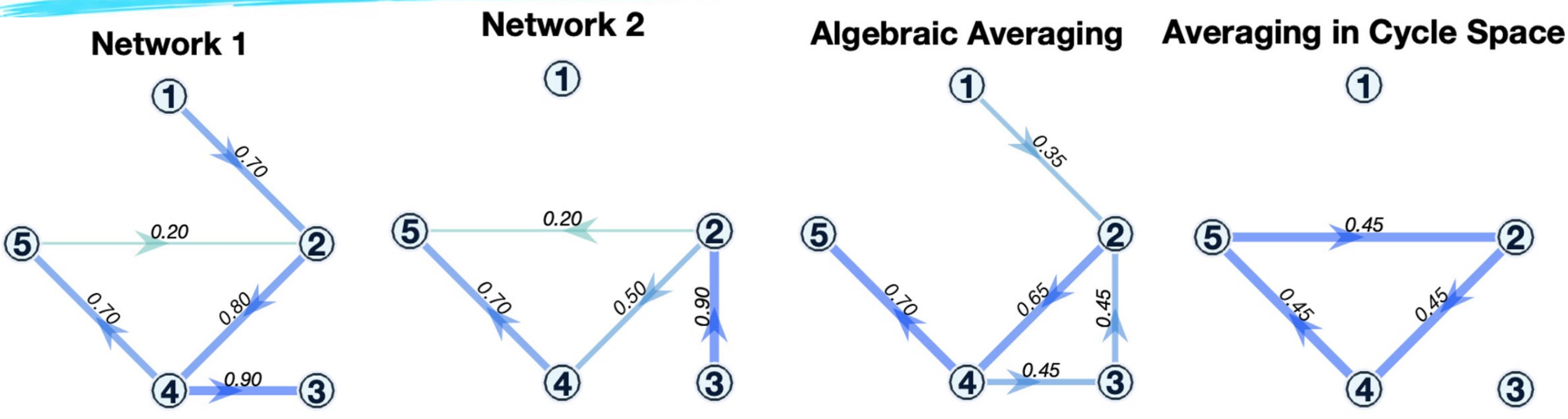
Harmonic subspace

$$H_k = \ker(\Delta_k) = \ker(B_k) \cap \ker(B_{k+1}^{\mathrm{T}})$$



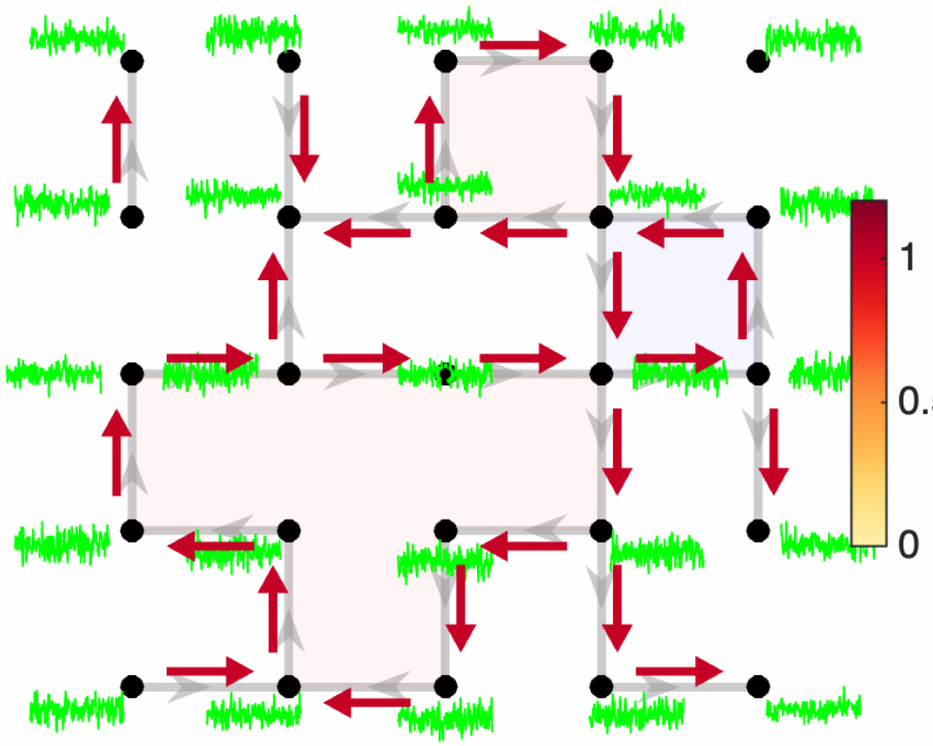
Harmonic cycles exclude boundary cycles. They represent topologically nontrivial loops (noncontractible).

$\ker(\Delta_1)$: Vector space of cycles



Cycles $C_1, C_2 \in \ker(\Delta_1) \dashrightarrow \alpha_1 C_1 + \alpha_2 C_2 \in \ker(\Delta_1)$

Diffusion along edges



Diffusion equation on edges

$$\dot{X}(t) = -\Delta_1 X(t)$$

↓ *Solution*

$$X(t) = e^{-\Delta_1 t} X(0)$$

↓ *Spectral expansion*

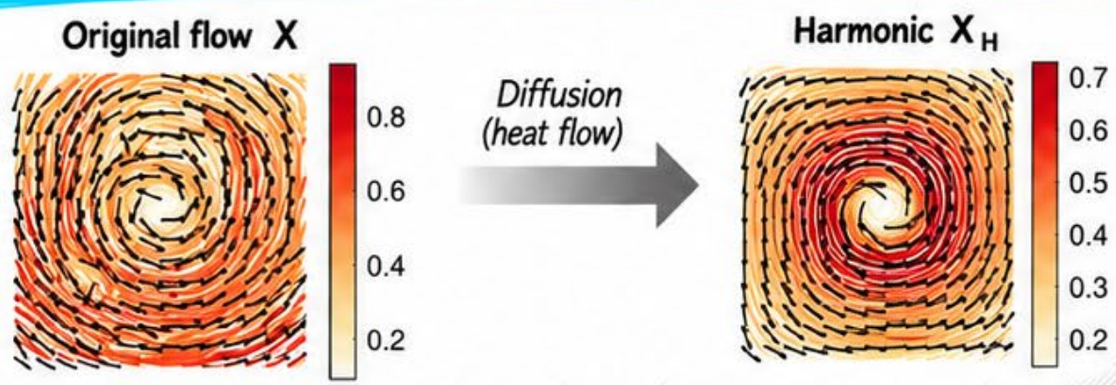
$$X(t) = \sum_k e^{-\lambda_k t} \psi_k \psi_k^\top X(0)$$

↓ $t \to \infty$

Harmonic cycle

$$X_H = \mathcal{P}_H X(0)$$

Intuition: diffusion removes non-harmonic flows

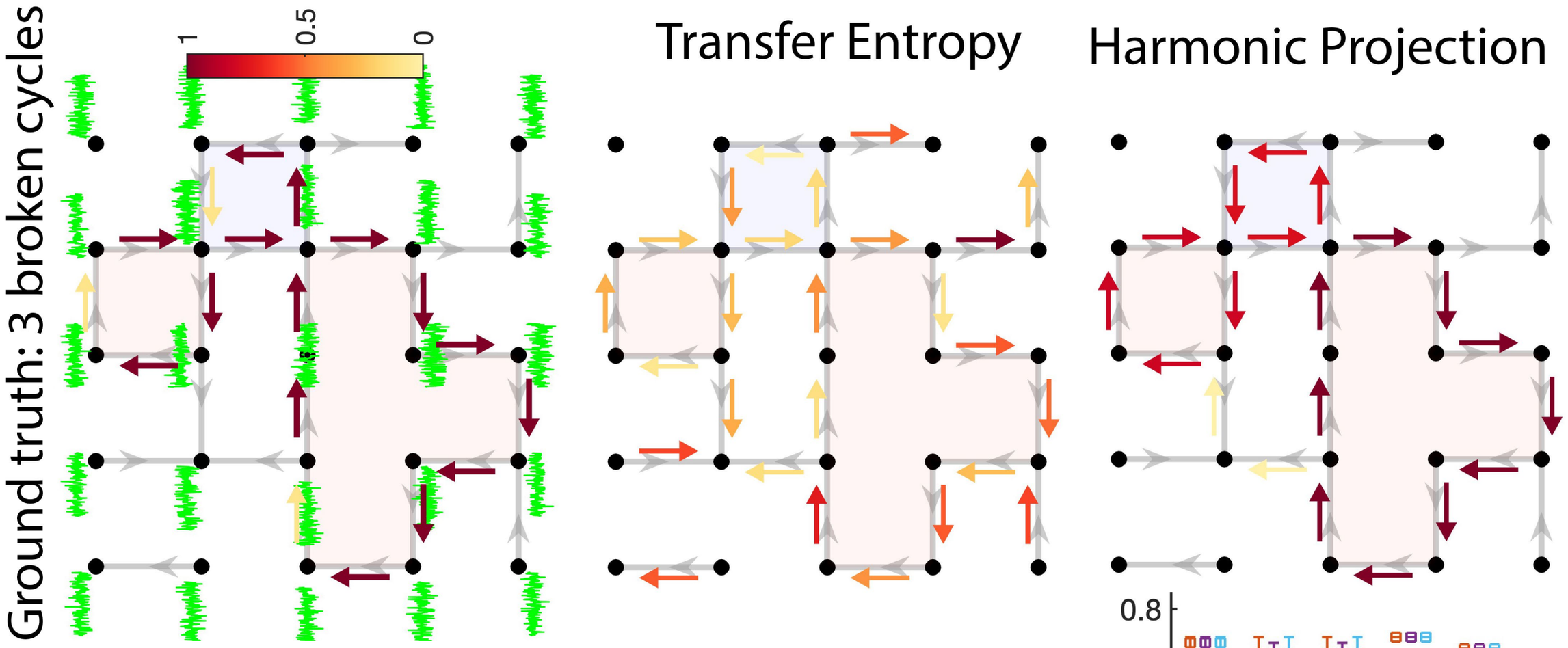


MATLAB code (Hodge 1-Laplacian)

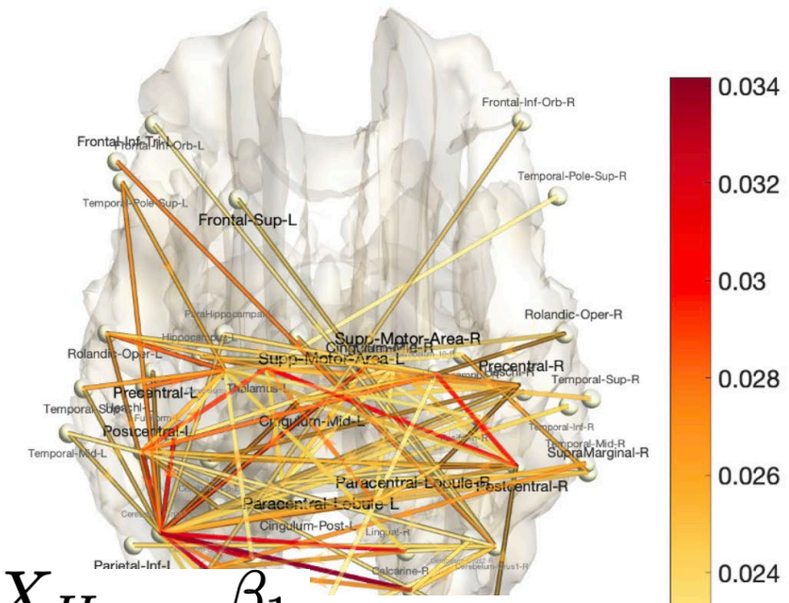
```
[Phi, Lambda] =
eig(full(Hodge1));
lambda = diag(Lambda);
c = Phi' * X0;
t = 2.0;
Xt = Phi * (exp(-t*lambda) .*
c);
```

Chung et al., 2026

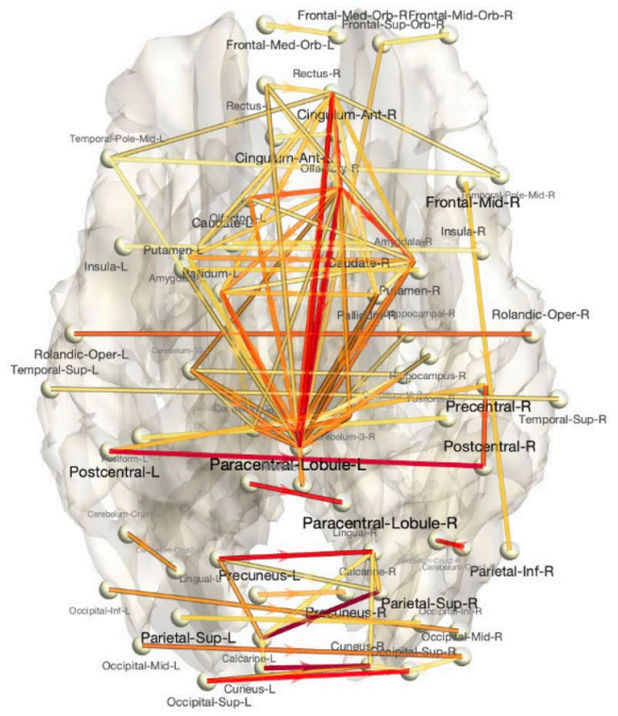
Performance in a cycle recovery task



$\bar{X}$

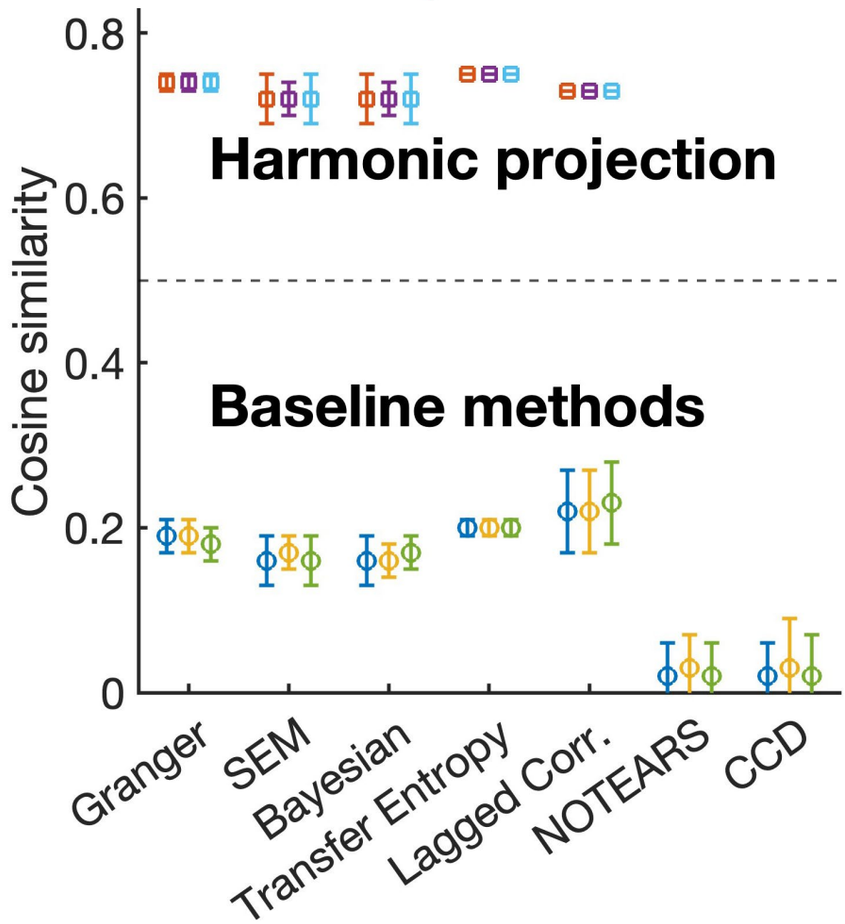


$\bar{X}_H$





$$\frac{\mathrm{Var}X_H}{\mathrm{Var}X} = \frac{\beta_1}{|E|} = 0.0801$$



Thank you



Any question, suggestion or collaboration

mkchung@wisc.edu